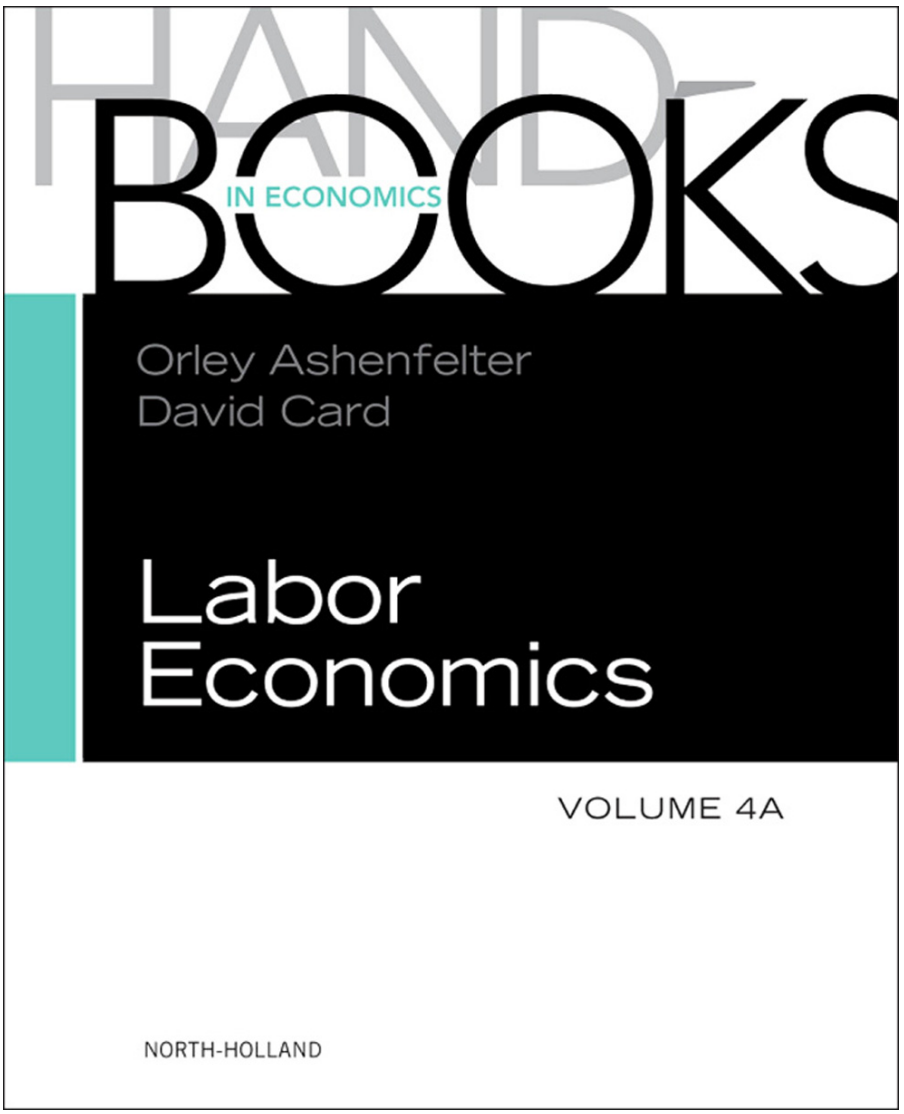

HANDBOOKS
IN ECONOMICS
Orley Ashenfelter
David Card
Labor Economics
VOLUME 4A
NORTH-HOLLAND

Handbook of
LABOR ECONOMICS

VOLUME *4A*

Edited by

ORLEY ASHENFELTER

DAVID CARD

2011

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

North Holland is an imprint of Elsevier

ELSEVIER

CONTENTS OF VOLUME 4A

CONTRIBUTORS TO VOLUME 4A

Nicole Fortin
UBC and CIFAR

Thomas Lemieux
UBC and NBER

Sergio Firpo
EESP-FGV and IZA

John A. List
Department of Economics, University of Chicago, 1126 East 59th Street, Chicago, IL 60637, USA

Imran Rasul
Department of Economics, University College London, Drayton House, 30 Gordon Street, London WC1E 6BT, United Kingdom

Gary Charness
University of California, Santa Barbara, United States

Peter Kuhn
University of California, Santa Barbara, United States

Michael P. Keane
University of Technology, Sydney and Arizona State University

Petra E. Todd
University of Pennsylvania

Kenneth I. Wolpin
University of Pennsylvania

John DiNardo
University of Michigan and NBER

David S. Lee
Princeton University and NBER

Eric French
Federal Reserve Bank of Chicago

Christopher Taber
Department of Economics, University of Wisconsin-Madison and NBER

Richard Rogerson
Arizona State University, United States

Robert Shimer
University of Chicago, United States

James B. Rebitzer
Boston University and NBER

Lowell J. Taylor
Carnegie Mellon University

CHAPTER 1

Decomposition Methods in Economics

Nicole Fortin [*] , **Thomas Lemieux** [**], **Sergio Firpo** [***]

[*] UBC and CIFAR
[**] UBC and NBER
[***] EESP-FGV and IZA

Contents

Handbook of Labor Economics, Volume 4a
ISSN 0169-7218, DOI 10.1016/S0169-7218(11)00407-2

Abstract

This chapter provides a comprehensive overview of decomposition methods that have been developed since the seminal work of Oaxaca and Blinder in the early 1970s. These methods are used to decompose the difference in a distributional statistic between two groups, or its change over time, into various explanatory factors. While the original work of Oaxaca and Blinder considered the case of the mean, our main focus is on other distributional statistics besides the mean, such as quantiles, the Gini coefficient or the variance. We discuss the assumptions required for identifying the different elements of the decomposition, as well as various estimation methods proposed in the literature. We also illustrate how these methods work in practice by discussing existing applications and working through a set of empirical examples throughout the paper.

JEL classification: J31; J71; C14; C21

Keywords: Decomposition; Counterfactual distribution; Inequality; Wage structure; Wage differentials; Discrimination

1. INTRODUCTION

What are the most important explanations accounting for pay differences between men and women? To what extent has wage inequality increased in the United States between 1980 and 2010 because of increasing returns to skill? Which factors are behind most of the growth in US GDP over the last 100 years? These important questions all share a common feature. They are typically answered using decomposition methods. The growth accounting approach pioneered by Solow (1957) and others is an early example of a decomposition approach aimed at quantifying the contribution of labor, capital, and unexplained factors (productivity) to US growth.[1] But it is in labor economics, starting with the seminal papers of Oaxaca (1973) and Blinder (1973), that decomposition methods have been used the most extensively. These two papers are among the most heavily cited in labor economics, and the Oaxaca-Blinder (OB) decomposition is now a standard tool in the toolkit of applied economists. A large number of methodological papers aimed at refining the OB decomposition, and expanding it to the case of distributional parameters besides the mean, have also been written over the past three decades.

The twin goals of this chapter are to provide a comprehensive overview of decomposition methods that have been developed since the seminal work of Oaxaca and

[1] See also Kendrick (1961), Denison (1962), and Jorgenson and Griliches (1967).

Blinder, and to suggest a list of best practices for researchers interested in applying these methods.[2] We also illustrate how these methods work in practice by discussing existing applications and working through a set of empirical examples throughout the chapter.

At the outset, it is important to note a number of limitations to decomposition methods that are beyond the scope of this chapter. As the above examples show, the goal of decomposition methods are often quite ambitious, which means that strong assumptions typically underlie these types of exercises. In particular, decomposition methods inherently follow a partial equilibrium approach. Take, for instance, the question "what would happen to average wages in the absence of unions?"As H. Gregg Lewis pointed out a long time ago (Lewis, 1963, 1986), there are many reasons to believe that eliminating unions would change not only the wages of union workers, but also those of non-union workers. In this setting, the observed wage structure in the non-union sector would not represent a proper counterfactual for the wages observed in the absence of unions. We discuss these general equilibrium considerations in more detail towards the end of the paper, but generally follow the standard partial equilibrium approach where observed outcomes for one group (or region/time period) can be used to construct various counterfactual scenarios for the other group.

A second important limitation is that while decompositions are useful for quantifying the contribution of various factors to a difference or change in outcomes in an accounting sense, they may not necessarily deepen our understanding of the mechanisms underlying the relationship between factors and outcomes. In that sense, decomposition methods, just like program evaluation methods, do not seek to recover behavioral relationships or "deep" structural parameters. By indicating which factors are quantitatively important and which are not, however, decompositions provide useful indications of particular hypotheses or explanations to be explored in more detail. For example, if a decomposition indicates that differences in occupational affiliation account for a large fraction of the gender wage gap, this suggests exploring in more detail how men and women choose their fields of study and occupations.

Another common use of decompositions is to provide some "bottom line" numbers showing the quantitative importance of particular empirical estimates obtained in a study. For example, while studies after studies show large and statistically significant returns to education, formal decompositions indicate that only a small fraction of US growth, or cross-country differences, in GDP per capita can be accounted for by changes or differences in educational achievement.

[2] We limit our discussion to so-called "regression-based" decomposition methods, where the decomposition focuses on explanatory factors, rather than decomposition methods that apply to additively decomposable indices, where the decomposition pertains to population sub-groups. Bourguignon and Ferreira (2005) and Bourguignon et al. (2008) are recent surveys discussing these methods.

Main themes and road map to the chapter

The original method proposed by Oaxaca and Blinder for decomposing changes or differences in the mean of an outcome variable has been considerably improved and expanded upon over the years. Arguably, the most important development has been to extend decomposition methods to distributional parameters other than the mean. For instance, Freeman (1980, 1984) went beyond a simple decomposition of the difference in mean wages between the union and non-union sector to look at the difference in the variance of wages between the two sectors.

But it is the dramatic increase in wage inequality observed in the United States and several other countries since the late 1970s that has been the main driving force behind the development of a new set of decomposition methods. In particular, the new methods introduced by Juhn et al. (1993) and DiNardo et al. (1996) were directly motivated by an attempt at better understanding the underlying factors behind inequality growth. Going beyond the mean introduces a number of important econometric challenges and is still an active area of research. As a result, we spend a significant portion of the chapter on these issues.

A second important development has been to use various tools from the program evaluation literature to (i) clarify the assumptions underneath popular decomposition methods, (ii) propose estimators for some of the elements of the decomposition, and (iii) obtain formal results on the statistical properties of the various decomposition terms. As we explain below, the key connection with the treatment effects literature is that the "unexplained" component of a Oaxaca decomposition can be interpreted as a treatment effect. Note that, despite the interesting parallel with the program evaluation literature, we explain in the paper that we cannot generally give a "causal" interpretation to the decomposition results.

The chapter also covers a number of other practical issues that often arise when working with decomposition methods. Those include the well known omitted group problem (Oaxaca and Ransom, 1999), and how to deal with cases where we suspect the true regression equation not to be linear.

Before getting into the details of the chapter, we provide here an overview of our main contributions by relating them to the original OB decomposition for the difference in mean outcomes for two groups A and B. The standard assumption used in these decompositions is that the outcome variable Y is linearly related to the covariates, X, and that the error term υ is conditionally independent of X:

$$Y_{gi} = \beta_{g0} + \sum_{k=1}^{K} X_{ik}\beta_{gk} + \upsilon_{gi}, \quad g = A, B, \tag{1}$$

where $\mathbb{E}(\upsilon_{gi}|X_i) = 0$, and X is the vector of covariates ($X_i = [X_{i1}, \ldots, X_{iK}]$). As is well known, the overall difference in average outcomes between group B and A,

$$\widehat{\Delta}_O^{\mu} = \overline{Y}_B - \overline{Y}_A,$$

can be written as:[3]

$$\widehat{\Delta}_O^{\mu} = \underbrace{(\widehat{\beta}_{B0} - \widehat{\beta}_{A0}) + \sum_{k=1}^{K} \overline{X}_{Bk} \left(\widehat{\beta}_{Bk} - \widehat{\beta}_{Ak}\right)}_{\widehat{\Delta}_S^{\mu} \text{ (Unexplained)}} + \underbrace{\sum_{k=1}^{K} \left(\overline{X}_{Bk} - \overline{X}_{Ak}\right) \widehat{\beta}_{Ak}}_{\widehat{\Delta}_X^{\mu} \text{ (Explained)}}$$

where $\widehat{\beta}_{g0}$ and $\widehat{\beta}_{gk}$ ($k = 1, \ldots, K$) are the estimated intercept and slope coefficients, respectively, of the regression models for groups $g = A, B$. The first term in the equation is what is usually called the "unexplained" effect in Oaxaca decompositions. Since we mostly focus on wage decompositions in this chapter, we typically refer to this first element as the "wage structure" effect (Δ_S^{μ}). The second component, Δ_X^{μ}, is a composition effect, which is also called the "explained" effect (by differences in covariates) in OB decompositions.

In the above decomposition, it is straightforward to compute both the overall composition and wage structure effects, and the contribution of each covariate to these two effects. Following the existing literature on decompositions, we refer to the overall decomposition (separating Δ_O^{μ} in its two components Δ_S^{μ} and Δ_X^{μ}) as an *aggregate decomposition*. The *detailed decomposition* involves subdividing both Δ_S^{μ}, the wage structure effect, and Δ_X^{μ}, the composition effect, into the respective contributions of each covariate, $\Delta_{S,k}^{\mu}$ and $\Delta_{X,k}^{\mu}$, for $k = 1, \ldots, K$.

The chapter is organized around the following "take away" messages:

[3] The decomposition can also be written by exchanging the reference group used for the wage structure and composition effects as follows:

$$\widehat{\Delta}_O^{\mu} = \left\{(\widehat{\beta}_{B0} - \widehat{\beta}_{A0}) + \sum_{k=1}^{K} \overline{X}_{Ak} \left(\widehat{\beta}_{Bk} - \widehat{\beta}_{Ak}\right)\right\} + \left\{\sum_{k=1}^{K} \left(\overline{X}_{Bk} - \overline{X}_{Ak}\right) \widehat{\beta}_{Bk}\right\}.$$

Alternatively, the so-called three-fold decomposition uses the same reference group for both effects, but introduces a third interaction term: $\widehat{\Delta}_O^{\mu} = \left\{(\widehat{\beta}_{B0} - \widehat{\beta}_{A0}) + \sum_{k=1}^{K} \overline{X}_{Ak} \left(\widehat{\beta}_{Bk} - \widehat{\beta}_{Ak}\right)\right\} + \left\{\sum_{k=1}^{K} \left(\overline{X}_{Bk} - \overline{X}_{Ak}\right) \widehat{\beta}_{Ak}\right\} + \left\{\sum_{k=1}^{K} \left(\overline{X}_{Bk} - \overline{X}_{Ak}\right) \left(\widehat{\beta}_{Bk} - \widehat{\beta}_{Ak}\right)\right\}$. While these various versions of the basic decomposition are used in the literature, using one or the other does not involve any specific estimation issues. For the sake of simplicity, we thus focus on the one decomposition introduced in the text for most of the chapter.

A. The wage structure effect can be interpreted as a treatment effect

This point is easily seen in the case where group B consists of union workers, and group A consists of non-union workers. The raw wage gap Δ^{μ} can be decomposed as the sum of the "effect" of unions on union workers, Δ_S^{μ}, and the composition effect linked to differences in covariates between union and non-union workers, Δ_X^{μ}. We can think of the effect of unions for each worker ($Y_{Bi} - Y_{Ai}$) as the individual treatment effect, while Δ_S^{μ} is the Average Treatment effect on the Treated (ATT). One difference between the program evaluation and decomposition approaches is that the composition effect Δ_X^{μ} is a key component of interest in a decomposition, while it is a selection bias resulting from a confounding factor to be controlled for in the program evaluation literature. By construction, however, one can obtain the composition effect from the estimated treatment effect since $ATT = \Delta_S^{\mu}$ and $\Delta_X^{\mu} = \Delta_O^{\mu} - \Delta_S^{\mu}$.

Beyond semantics, there are a number of advantages associated with representing the decomposition component Δ_S^{μ} as a treatment effect:

- The zero conditional mean assumption ($\mathbb{E}(\upsilon|X) = 0$) usually invoked in OB decompositions (as above) is not required for consistently estimating the ATT (or Δ_S^{μ}). The mean independence assumption can be replaced by a weaker ignorability assumption. Under ignorability, unobservables do not need to be independent (or mean independent) of X as long as their conditional distribution given X is the same in groups A and B. In looser terms, this "selection based on observables" assumption allows for selection biases as long they are the same for the two groups. For example, if unobservable ability and education are correlated, a linear regression of Y on X will not yield consistent estimates of the structural parameters (i.e. the return to education). But the aggregate decomposition remains valid as long as the dependence structure between ability and education is the same in group A and B.
- A number of estimators for the ATT have been proposed in the program evaluation literature including Inverse Probability Weighting (IPW), matching and regression methods. Under ignorability, these estimators are consistent for the ATT (or Δ_S^{μ}) even if the relationship between Y and X is not linear. The statistical properties of these non-parametric estimators are also relatively well established. For example, Hirano et al. (2003) show that IPW estimators of the ATT are efficient. Firpo (2007) similarly shows that IPW is efficient for estimating quantile treatment effects. Accordingly, we can use the results from the program evaluation literature to show that decomposition methods based on reweighting techniques are efficient for performing decompositions.[4]

[4] Firpo (2010) shows that for any smooth functional of the reweighted cdf, efficiency is achieved. In other words, decomposing standard distributional statistics such as the variance, the Gini coefficient, or the interquartile range using the reweighting method suggested by DiNardo et al. (1996) will be efficient. Note, however, that this result does not apply to the (more complicated) case of the density considered by DiNardo et al. (1996) where non-parametric estimation is involved.

- When the distribution of covariates is different across groups, the *ATT* depends on the characteristics of group *B* (unless there is no heterogeneity in the treatment effect, i.e. $\beta_{Bk} = \beta_{Ak}$ for all k). The subcomponents of Δ_S^μ associated with each covariate k, $\overline{X}_{Bk}\,(\beta_{Bk} - \beta_{Ak})$, can be (loosely) interpreted as the "contribution" of the covariate k to the *ATT*. This helps understand the issues linked to the well-known "omitted group problem" in OB decompositions (see, for example Oaxaca and Ransom, 1999).

B. Going beyond the mean is a "solved" problem for the aggregate decomposition

As discussed above, estimation methods from the program evaluation literature can be directly applied for performing an aggregate decomposition of the gap Δ_O^μ into its two components Δ_S^μ and Δ_X^μ. While most of the results in the program evaluation literature have been obtained in the case of the mean (e.g., Hirano et al., 2003), they can also be extended to the case of quantiles (Firpo, 2007) or more general distribution parameters (Firpo, 2010). The *IPW* estimator originally proposed in the decomposition literature by DiNardo et al. (1996) or matching methods can be used to perform the decomposition under the assumption of ignorability. More parametric approaches such as those proposed by Juhn et al. (1993), Donald et al. (2000) and Machado and Mata (2005) could also be used. These methods involve, however, a number of assumptions and/or computational difficulties that can be avoided when the sole goal of the exercise is to perform an aggregate decomposition. By contrast, *IPW* methods involve no parametric assumptions and are an efficient way of estimating the aggregate decomposition.

It may be somewhat of an overstatement to say that computing the aggregate decomposition is a "solved" problem since there is still ongoing research on the small sample properties of various treatment effect estimators (see, for example, Busso et al., 2009). Nonetheless, performing an aggregate decomposition is relatively straightforward since several easily implementable estimators with good asymptotics properties are available.

C. Going beyond the mean is more difficult for the detailed decomposition

Until recently, no comprehensive approach was available for computing a detailed decomposition of the effect of single covariates for a distributional statistic ν other than the mean. One popular approach for estimating the subcomponents of Δ_S^ν is Machado and Mata (2005)'s method, which relies on quantile regressions for each possible quantile, combined with a simulation procedure. For the subcomponents of Δ_X^ν, DiNardo et al. (1996) suggest a reweighting procedure to compute the contribution of a dummy covariate (like union status) to the aggregate composition effect Δ_X^ν. Altonji et al. (2008) implemented a generalization of this approach to the case of either continuous or categorical covariates. Note, however, that these latter methods are generally *path dependent*, that is, the decomposition results depend on the order in

which the decomposition is performed. Later in this chapter, we show how to make the contribution of the last single covariate path independent in the spirit of Gelbach (2009).

One comprehensive approach, very close in spirit to the original OB decomposition, which is *path independent*, uses the recentered influence function (**RIF**) regressions recently proposed by Firpo et al. (2009). The idea is to use the (recentered) influence function for the distribution statistic of interest instead of the usual outcome variable Y as the left hand side variable in a regression. In the special case of the mean, the recentered influence function is Y, and a standard regression is estimated, as in the case of the OB decomposition.

More generally, once the **RIF** regression has been estimated, the estimated coefficients can be used to perform the detailed decomposition in the same way as in the standard OB decomposition. The downside of this approach is that **RIF** regression coefficients only provide a local approximation for the effect of changes in the distribution of a covariate on the distributional statistics of interest. The question of how accurate this approximation is depends on the application at hand.

D. The analogy between quantile and standard (mean) regressions is not helpful

If the mean can be decomposed using standard regressions, can we also decompose quantiles using simple quantile regressions? Unfortunately, the answer is negative. The analogy with the case of the mean just does not apply in the case of quantile regressions.

To understand this point, it is important to recall that the coefficient β in a standard regression has two distinct interpretations. Under the *conditional mean interpretation*, β indicates the effect of X on the conditional mean $\mathbb{E}(Y|X)$ in the model $\mathbb{E}(Y|X) = X\beta$. Using the law of iterated expectations, we also have $\mathbb{E}(Y) = \mathbb{E}_X[\mathbb{E}(Y|X)] = \mathbb{E}(X)\beta$. This yields an *unconditional mean interpretation* where β can be interpreted as the effect of increasing the mean value of X on the (unconditional) mean value of Y. It is this particular property of regression models, and this particular interpretation of β, which is used in OB decompositions.

By contrast, only the conditional quantile interpretation is valid in the case of quantile regressions. As we discuss in more detail later, a quantile regression model for the τth conditional quantile $Q_\tau(X)$ postulates that $Q_\tau(X) = X\beta_\tau$. By analogy with the case of the mean, β_τ can be interpreted as the effect of X on the τth conditional quantile of Y given X. The law of iterated expectations does not apply in the case of quantiles, so $Q_\tau \neq \mathbb{E}_X[Q_\tau(X)] = \mathbb{E}(X)\beta_\tau$, where Q_τ is the unconditional quantile. It follows that β_τ cannot be interpreted as the effect of increasing the mean value of X on the unconditional quantile Q_τ.

This greatly limits the usefulness of quantile regressions in decomposition problems. Machado and Mata (2005) suggest estimating quantile regressions for all $\tau \in [0, 1]$ as a way of characterizing the full conditional distribution of Y given X. The estimates are

then used to construct the different components of the aggregate decomposition using simulation methods. Compared to other decomposition methods, one disadvantage of this method is that it is computational intensive.

An alternative regression approach where the estimated coefficient can be interpreted as the effect of increasing the mean value of X on the unconditional quantile Q_τ (or other distributional parameters) has recently been proposed by Firpo et al. (2009). As we mention above, this method provides one of the few options available for computing a detailed decomposition for distributional parameters other than the mean.

E. Decomposing proportions is easier than decomposing quantiles

A cumulative distribution provides a one-to-one mapping between (unconditional) quantiles and the proportion of observations below this quantile. Performing a decomposition on proportions is a fairly standard problem. One can either run a linear probability model and perform a traditional OB decomposition, or do a non-linear version of the decomposition using a logit or probit model.

Decompositions of quantiles can then be obtained by inverting back proportions into quantiles. Firpo et al. (2007) propose doing so using a first order approximation where the elements of the decomposition for a proportion are transformed into elements of the decomposition for the corresponding quantile by dividing by the density (slope of the cumulative distribution function). This can be implemented in practice by estimating recentered influence function (RIF) regressions (see Firpo et al., 2009).

A related approach is to decompose proportions at every point of the distribution (e.g. at each percentile) and invert back the whole fitted relationship to quantiles. This can be implemented in practice using the distribution regression approach of Chernozhukov et al. (2009).

F. There is no general solution to the "omitted group" problem

As pointed out by Jones (1983) and Oaxaca and Ransom (1999) among others, in the case of categorical covariates, the various elements of Δ_S^μ in a detailed decomposition arbitrarily depend on the choice of the omitted group in the regression model. In fact, this interpretation problem may arise for any covariate, including continuous covariates, that does not have a clearly interpretable baseline value. This problem has been called an identification problem in the literature (Oaxaca and Ransom, 1999; Yun, 2005). But as pointed out by Gelbach (2002), it is better viewed as a conceptual problem with the detailed part of the decomposition for the wage structure effect.

As discussed above, the effect $\beta_{B0} - \beta_{A0}$ for the omitted group can be interpreted as an average treatment effect among the omitted group (group for which $X_k = 0$ for all $k = 1, \ldots, K$). The decomposition then corresponds to a number of counterfactual experiments asking "by how much the treatment effect would change if X_k was switched from its value in the omitted group (0) to its average value ($\overline{X}_{Bk}$)"? In cases like the

gender wage gap where the treatment effect analogy is not as clear, the same logic applied, nonetheless. For example, one could ask instead "by how much the average gender gap would change if actual experience (X_k) was switched from its value in the omitted group (0) to its average value ($\overline{X}_{Bk}$)?"

Since the choice of the omitted group is arbitrary, the elements of the detailed decomposition can be viewed as arbitrary as well. In cases where the omitted group has a particular economic meaning, the elements of the detailed decomposition are more interpretable as they correspond to interesting counterfactual exercises. In other cases the elements of the detailed decomposition are not economically interpretable. As a result, we argue that attempts at providing a general "solution" to the omitted group problem are misguided. We discuss instead the importance of using economic reasoning to propose some counterfactual exercise of interest, and suggest simple techniques to easily compute these counterfactual exercises for any distributional statistics, and not only the mean.

Organization of the chapter

The different methods covered in the chapter, along with their key assumptions and properties are listed in Table 1. The list includes an example of one representative study for each method, focusing mainly on studies on the gender and racial gap (see also Altonji and Blank, 1999), to facilitate comparison across methods. A detailed discussion of the assumptions and properties follows in the next section. The mean decomposition methodologies comprise the classic OB decomposition, as well as extensions that appeal to complex counterfactuals and that apply to limited dependent variable models. The methodologies that go beyond the mean include the classic variance decomposition, methods based on residual imputation, methods based on conditional quantiles and on estimating the conditional distribution, and methods based on reweighting and RIF-regressions.

Since there are a number of econometric issues involved in decomposition exercises, we start in Section 2 by establishing what are the parameters of interest, their interpretation, and the conditions for identification in decomposition methods. We also introduce a general notation that we use throughout the chapter. Section 3 discusses exhaustively the case of decomposition of differences in means, as originally introduced by Oaxaca (1973) and Blinder (1973). This section also covers a number of ongoing issues linked to the interpretation and estimation of these decompositions. We then discuss decompositions for distributional statistics other than the mean in Sections 4 and 5. Section 4 looks at the case of the aggregate decomposition, while Section 5 focuses on the case of the detailed decomposition. Finally, we discuss a number of limitations and extensions to these standard decomposition methods in Section 6. Throughout the chapter, we illustrate the "nuts and bolts" of decomposition methods using empirical examples, and discuss important applications of these methods in the applied literature.

Table 1 Maintained assumptions and properties of major decomposition methodologies.

Methods	Assumptions[a]	Properties: limitations and advantages	Representative applications
Mean decomposition			
3.1. Standard OB: Oaxaca (1973)-Blinder (1973)	Linearity of $E[Y\|X]$, Zero conditional mean	Path independent, Detailed decomposition	Gender and racial wage gaps: O'Neill and O'Neill (2006)
3.3. Weighted or pooled OB: Oaxaca and Ransom (1994), Cotton (1998)	Complex counterfactual, Linearity of $E[Y\|X]$, Zero conditional mean	Path independent, Detailed decomposition	Racial/Ethnic wage gaps: Reimers (1983)
3.5 Non-linear OB: Fairlie (2005), Bauer and Sinning (2008)	Non-Linearity of $E[Y\|X]$	Path dependent, Detailed decomposition	Racial gap in self-employment: Fairlie (1999)
Going beyond the mean			
4.1. Variance decompositions	Linearity of $V(Y\|X)$, Invariance of conditional variance	No detailed decomposition	Union wage differentials: Freeman (1980, 1984)
4.3. Residual imputation procedure: Juhn et al. (1991, 1993)	Linearity of $E[Y\|X]$, Conditional rank preservation, Complex counterfactual[b]	No detailed decomposition	Gender gap across countries: Blau and Kahn (1992)
4.4. Quantile regressions methods: Machado and Mata (2005), Chernozhukov et al. (2009)	Linearity of $Q_\tau(Y\|X)$, Conditional rank preservation	No detailed decomposition	Gender glass ceiling: Albrecht et al. (2003)

(*continued on next page*)

Table 1 (continued)

Methods	Assumptions[a]	Properties: limitations and advantages	Representative applications
4.5 Inverse propensity reweighting: DiNardo et al. (1996)	Invariance of conditional distribution	Path dependent	Immigrant/Resident wage differentials: Chiquiar and Hanson (2005)
4.6 Estimation of conditional distribution: Chernozhukov et al. (2009)	Invariance of conditional distribution, Conditional rank preservation	Path dependent	Racial wage gap: Melly (2006)
5.2 RIF regressions: Firpo et al. (2007, 2009)	Invariance of conditional distribution	Path independent, Detailed decomposition	Racial wage gap: Heywood and Parent (2009)

[a] Unless otherwise indicated, the different methodologies appeal to a simple counterfactual treatment.

[b] In some applications, the counterfactual is an average over time periods or over countries.

2. IDENTIFICATION: WHAT CAN WE ESTIMATE USING DECOMPOSITION METHODS?

As we will see in subsequent sections, a large and growing number of procedures are available for performing decompositions of the mean or more general distributional statistics. But despite this rich literature, it is not always clear what these procedures seek to estimate, and what conditions need to be imposed to recover the underlying objects of interest. The main contribution of this section is to provide a more formal theory of decompositions where we clearly define what it is that we want to estimate using decompositions, and what are the assumptions required to identify the population parameters of interest. In the first part of the section, we discuss the case of the aggregate decomposition. Since the estimation of the aggregate decomposition is closely related to the estimation of treatment effects (see the introduction), we borrow heavily from the identification framework used in the treatment effects literature. We then move to the case of the detailed decomposition, where additional assumptions need to be introduced to identify the parameters of interest. We end the section by discussing the connection between program evaluation and decompositions, as well as the more general issue of causality in this context.

Decompositions are often viewed as simple accounting exercises based on correlations. As such, results from decomposition exercises are believed to suffer from the same shortcomings as OLS estimates, which cannot be interpreted as valid estimates of some underlying causal parameters in most circumstances. The interpretation of what decomposition results mean becomes even more complicated in the presence of general equilibrium effects.

In this section, we argue that these interpretation problems are linked in part to the lack of a formal identification theory for decompositions. In econometrics, the standard approach is to first discuss identification (what we want to estimate, and what assumptions are required to interpret these estimates as sample counterparts of parameters of interest) and then introduce estimation procedures to recover the object we want to identify. In the decomposition literature, most papers jump directly to the estimation issues (i.e. discuss procedures) without first addressing the identification problem.[5]

To simplify the exposition, we use the terminology of labor economics, where, in most cases, the agents are workers and the outcome of interest is wages. Decomposition methods can also be applied in a variety of other settings, such as gaps in test scores between gender (Sohn, 2008), schools (Krieg and Storer, 2006) or countries (McEwan and Marshall, 2004).

[5] One possible explanation for the lack of discussion of identification assumptions is that they were reasonably obvious in the case of the original OB decompositions for the mean. The situation is quite a bit more complex, however, in the case of distributional statistics other than the mean. Note also that some recent papers have started addressing these identification issues in more detail. See, for instance, Firpo et al. (2007), and Chernozhukov et al. (2009).

Throughout the chapter, we restrict our discussion to the case of a decomposition for two mutually exclusive groups. This rules out decomposing wage differentials between overlapping groups like Blacks, Whites, and Hispanics, who can be Black or White.[6] In this setting, the dummy variable method (Cain, 1986) with interactions is a more natural way of approaching the problem. Then one can use Gelbach (2009)'s approach, which appeals to the omitted variables bias formula, to compute a detailed decomposition.

The assumption of mutually exclusive groups is not very restrictive, however, since most decomposition exercises fall into this category:

Assumption 1 (*Mutually Exclusive Groups*). The population of agents can be divided into two mutually exclusive groups, denoted A and B. Thus, for an agent i, $D_{Ai} + D_{Bi} = 1$, where $D_{gi} = \mathbb{1}\{i \text{ is in } g\}$, $g = A, B$, and $\mathbb{1}\{\cdot\}$ is the indicator function.

We are interested in comparing features of the wage distribution for two groups of workers: A and B. We observe wage Y_i for worker i, which can be written as $Y_i = D_{gi}Y_{gi}$, for $g = A, B$, where Y_{gi} is the wage worker i would receive in group g. Obviously, if worker i belongs to group A, for example, we only observe Y_{Ai}.

As in the treatment effects literature, Y_{Ai} and Y_{Bi} can be interpreted as two potential outcomes for worker i. While we only observe Y_{Ai} when $D_{Ai} = 1$, and Y_{Bi} when $D_{Bi} = 1$, decompositions critically rely on counterfactual exercises such as "what would be the distribution of Y_A for workers in group B ?". Since we do not observe this counterfactual wage $Y_{A|D_B}$ for these workers, some assumptions are required for estimating this counterfactual distribution.

2.1. Case 1: The aggregate decomposition

2.1.1. The overall wage gap and the structural form

Our identification results for the aggregate decomposition are very general, and hold for any distributional statistic.[7] Accordingly, we focus on general distributional measures in this subsection of the chapter.

Consider the case where the distributional statistic of interest is $\nu(F_{Y_g|D_s})$, where $\nu: \mathcal{F}_\nu \to \mathbb{R}$ is a real-valued functional, and where $\mathcal{F}_\nu$ is a class of distribution functions such that $F_{Y_g|D_s} \in \mathcal{F}_\nu$ if $|\nu(F_{Y_g|D_s})| < \infty$, $g, s = A, B$. The distribution function $F_{Y_g|D_s}$ represents the distribution of the (potential) outcome Y_g for workers in group s. $F_{Y_g|D_s}$ is an observed distribution when $g = s$, and a counterfactual distribution when $g \neq s$.

[6] Alternatively, the overlapping issue can bypassed by excluding Hispanics from the Black and White groups.

[7] Many papers (DiNardo et al., 1996; Machado and Mata, 2005; Chernozhukov et al., 2009) have proposed methodologies to estimate and decompose entire distributions (or densities) of wages, but the decomposition results are ultimately quantified through the use of distributional statistics. Analyses of the entire distribution look at several of these distributional statistics simultaneously.

The overall ν-difference in wages between the two groups measured in terms of the distributional statistic ν is

$$\Delta_O^\nu = \nu\left(F_{Y_B|D_B}\right) - \nu\left(F_{Y_A|D_A}\right). \tag{2}$$

The more common distributional statistics used to study wage differentials are the mean and the median. The wage inequality literature has focused on the variance of log wages, the Gini and Theil coefficients, and the differentials between the 90th and 10th percentiles, the 90th and 50th percentiles, and the 50th and 10th percentiles. These latter measures provide a simply way of distinguishing what happens at the top and bottom end of the wage distribution. Which statistic ν is most appropriate depends on the problem at hand.

A typical aim of decomposition methods is to divide Δ_O^ν, the *ν-overall wage gap* between the two groups, into a component attributable to differences in the observed characteristics of workers, and a component attributable to differences in wage structures. In our setting, the wage structure is what links observed characteristics, as well as some unobserved characteristics, to wages.

The decomposition of the overall difference into these two components depends on the construction of a meaningful counterfactual wage distribution. For example, counterfactual states of the world can be constructed to simulate what the distribution of wages would look like if workers had different returns to observed characteristics. We may want to ask, for instance, what would happen if group A workers were paid like group B workers, or if women were paid like men? When the two groups represent different time periods, we may want to know what would happen if workers in year 2000 had the same characteristics as workers in 1980, but were still paid as in 2000. A more specific counterfactual could keep the return to education at its 1980 level, but set all the other components of the wage structure at their 2000 levels.

As these examples illustrate, counterfactuals used in decompositions often consist of manipulating structural wage setting functions (i.e. the wage structure) linking the observed and unobserved characteristics of workers to their wages for each group. We formalize the role of the wage structure using the following assumption:

Assumption 2 (*Structural Form*). A worker i belonging to either group A or B is paid according to the wage structure, m_A and m_B, which are functions of the worker's observable (X) and unobservable (ε) characteristics:

$$Y_{Ai} = m_A\left(X_i, \varepsilon_i\right) \quad \text{and} \quad Y_{Bi} = m_B\left(X_i, \varepsilon_i\right), \tag{3}$$

where ε_i has a conditional distribution $F_{\varepsilon|X}$ given X, and $g = A, B$.

While the wage setting functions are very general at this point, the assumption implies that there are only three reasons why the wage distribution can differ between

group A and B. The three potential sources of differences are (i) differences between the wage setting functions m_A and m_B, (ii) differences in the distribution of observable (X) characteristics, and (iii) differences in the distribution of unobservable (ε) characteristics. The aim of the aggregate decomposition is to separate the contribution of the first factor (differences between m_A and m_B) from the two others.

When the counterfactuals are based on the alternative wage structure (i.e. using the observed wage structure of group A as a counterfactual for group B), decompositions can easily be linked to the treatment effects literature. However, other counterfactuals may be based on hypothetical states of the world, that may involve *general equilibrium effects*. For example, we may want to ask what would be the distribution of wages if group A workers were paid according to the pay structure that would prevail if there were no B workers, for example if there were no union workers. Alternatively, we may want to ask what would happen if women were paid according to some non-discriminatory wage structure (which differs from what is observed for either men or women)?

We use the following assumption to restrict the analysis to the first type of counterfactuals.

Assumption 3 (*Simple Counterfactual Treatment*). A counterfactual wage structure, m^C, is said to correspond to a simple counterfactual treatment when it can be assumed that $m^C(\cdot,\cdot) \equiv m_A(\cdot,\cdot)$ for workers in group B, or $m^C(\cdot,\cdot) \equiv m_B(\cdot,\cdot)$ for workers in group A.

It is helpful to represent the assumption using the potential outcomes framework introduced earlier. Consider $Y_{g|D_s}$,where $g = A, B$ indicates the potential outcome, while $s = A, B$ indicates group membership. For group A, the observed wage is $Y_{A|D_A}$, while $Y^C_{B|D_A}$ represents the counterfactual wage. For group B, $Y_{B|D_B}$ is the observed wage while the counterfactual wage is $Y^C_{A|D_B}$. Note that we add the superscript C to highlight counterfactual wages. For instance, consider the case where workers in group B are unionized, while workers in group A are not unionized. The dichotomous variable D_B indicates the union status of workers. For a worker i in the union sector ($D_B = 1$), the observed wage under the "union" treatment is $Y_{B|D_B,i} = m_B(X_i, \varepsilon_i)$, while the counterfactual wage that would prevail if the worker was not unionized is $Y^C_{A|D_B,i} = m^C(X_i, \varepsilon_i) = m_A(X_i, \varepsilon_i)$, $i \in B$. An alternative counterfactual could ask what would be the wage of a non-union worker j if this worker was unionized $Y^C_{B|D_A,j} = m^C(X_j, \varepsilon_j) = m_B(X_j, \varepsilon_j)$, $j \in A$. We note that the choice of which counterfactual to choose is analogous to the choice of reference group in standard OB decomposition.[8]

What Assumption 3 rules out is the existence of another counterfactual wage structure such as $m^*(\cdot)$ that represents how workers would be paid if there were no unions in

[8] When we construct the counterfactual $Y^C_{g|D_s}$, we choose g to be the reference group and s the group whose wages are "adjusted". Thus counterfactual women's wages if they were paid like men would be $Y^C_{m|D_f}$, although the gender gap example is more difficult to conceive in the treatment effects literature.

the labor market. Unless there are no general equilibrium effects, we would expect that $m^*(\cdot) \neq m_A(\cdot)$, and, thus, Assumption 3 to be violated.

2.1.2. Four decomposition terms

With this setup in mind, we can now decompose the overall difference Δ_O^v into the four following components of interest:

D.1 Differences associated with the return to observable characteristics under the structural m functions. For example, one may have the following counterfactual in mind: What if everything but the return to X was the same for the two groups?

D.2 Differences associated with the return to unobservable characteristics under the structural m functions. For example, one may have the following counterfactual in mind: What if everything but the return to ε was the same for the two groups?

D.3 Differences in the distribution of observable characteristics. We have here the following counterfactual in mind: What if everything but the distribution of X was the same for the two groups?

D.4 Differences in the distribution of unobservable characteristics. We have the following counterfactual in mind: What if everything but the distribution of ε was the same for the two groups?

Obviously, because unobservable components are involved, we can only decompose Δ_O^v into the four decomposition terms after imposing some assumptions on the joint distribution of observable and unobservable characteristics. Also, unless we make additional separability assumptions on the structural forms represented by the m functions, it is virtually impossible to separate out the contribution of returns to observables from that of unobservables. The same problem prevails when one tries to perform a detailed decomposition in returns, that is, provide the contribution of the return to each covariate separately.

2.1.3. Imposing identification restrictions: overlapping support

The first assumption we make to simplify the discussion is to impose a common support assumption on the observables and unobservables. Further, this assumption ensures that no single value of $X = x$ or $\varepsilon = e$ can serve to identify membership into one of the groups.

Assumption 4 (*Overlapping Support*). Let the support of all wage setting factors $\left[X', \varepsilon'\right]'$ be $\mathcal{X} \times \mathcal{E}$. For all $\left[x', e'\right]'$ in $\mathcal{X} \times \mathcal{E}$, $0 < \Pr[D_B = 1 | X = x, \varepsilon = e] < 1$.

Note that the overlapping support assumption rules out cases where inputs may be different across the two wage setting functions. The case of the wage gap between

immigrant and native workers is an important example where the X vector may be different for two groups of workers. For instance, the wage of immigrants may depend on their country of origin and their age at arrival, two variables that are not defined for natives. Consider also the case of changes in the wage distribution over time. If group A consists of workers in 1980, and group B of workers in 2000, the difference in wages over time should take into account the fact that many occupations of 2000, especially those linked to information technologies, did not even exist in 1980. Thus, taking those differences explicitly into account could be important for understanding the evolution of the wage distribution over time.

The case with different inputs can be formalized as follows. Assume that for group A, there is a $d_A + l_A$ vector of observable and unobservable characteristics $[X'_A, \varepsilon'_A]'$ that may include components not included in the $d_B + l_B$ vector of characteristics $[X'_B, \varepsilon'_B]'$ for group B, where d_g and l_g denote the length of the X_g and ε_g vectors, respectively. Define the intersection of these characteristics by the $d + l$ vector $[X', \varepsilon']'$, which represent characteristics common to both groups. The respective complements, which are group-specific characteristics, are denoted by tilde as $[X'_{\tilde{A}}, \varepsilon'_{\tilde{A}}]'$and $[X'_{\tilde{B}}, \varepsilon'_{\tilde{B}}]'$, such that $[X'_{\tilde{A}}, \varepsilon'_{\tilde{A}}]' \cup [X', \varepsilon']' = [X'_A, \varepsilon'_A]'$ and $[X'_{\tilde{B}}, \varepsilon'_{\tilde{B}}]' \cup [X', \varepsilon']' = [X'_B, \varepsilon'_B]'$.

In that context, the overlapping support assumption could be restated by letting the support of all wage setting factors $[X'_A, \varepsilon'_A]' \cup [X'_B, \varepsilon'_B]'$ be $\mathcal{X} \times \mathcal{E}$. The overlapping support assumption would then guarantee that, for all $[x', e']'$ in $\mathcal{X} \times \mathcal{E}$, $0 < \Pr[D_B = 1 | [X'_A, X'_B] = x, [\varepsilon'_A, \varepsilon'_B] = e] < 1$. The assumption rules out the existence of the vectors $[X'_{\tilde{A}}, \varepsilon'_{\tilde{A}}]$ and $[X'_{\tilde{B}}, \varepsilon'_{\tilde{B}}]$.

In the decomposition of gender wage differentials, it is not uncommon to have explanatory variables for which this condition does not hold. Black et al. (2008) and Ñopo (2008) have proposed alternative decompositions based on matching methods to address cases where they are severe gaps in the common support assumption (for observables). For example, Ñopo (2008) divides the gap into four additive terms. The first two are analogous to the above composition and wage structure effects, but they are computed only over the common support of the distributions of observable characteristics, while the other two account for differences in support.

2.1.4. Imposing identification restrictions: ignorability

We cannot separate out the decomposition terms (D.1) and (D.2) unless we impose some separability assumptions on the functional forms of m_A and m_B. For highly complex nonlinear functions of observables X and unobservables ε, there is no clear definition of what would be the component of the m functions associated with either X or ε. For instance, if X and ε represent years of schooling and unobserved ability, respectively, we may expect the return to schooling to be higher for high ability workers. As a result,

there is an interaction term between X or ε in the wage equation $m(X, \varepsilon)$, which makes it hard to separate the contribution of these two variables to the wage gap.

Thus, consider the decomposition term D.1* that combines (D.1) and (D.2):

D.1* Differences associated with the return to observable and unobservable characteristics in the structural m functions.

This decomposition term solely reflects differences in the m functions. We call this decomposition term Δ_S^ν, or the "ν-wage structure effect" on the "ν-overall difference", Δ_O^ν. The key question here is how to identify the three decomposition terms (D.1*), (D.3) and (D.4) which, under Assumption 4, fully describe Δ_O^ν?

We denote the decomposition terms (D.3) and (D.4) as Δ_X^ν and Δ_ε^ν, respectively. They capture the impact of differences in the distributions of X and ε between groups B and A on the overall difference, Δ_O^ν. We can now write

$$\Delta_O^\nu = \Delta_S^\nu + \Delta_X^\nu + \Delta_\varepsilon^\nu.$$

Without further assumptions we still cannot identify these three terms. There are two problems. First, we have not imposed any assumption for the identification of the m functions, which could help in our identification quest. Second, we have not imposed any assumption on the distribution of unobservables. Thus, even if we fix the distribution of covariates X to be the same for the two groups, we cannot clearly separate all three components because we do not observe what would happen to the unobservables under this scenario.

Therefore, we need to introduce an assumption to make sure that the effect of manipulations of the distribution of observables X will not be confounded by changes in the distribution of ε. As we now show formally, the assumption required to rule out these confounding effects is the well-known ignorability, or unconfoundedness, assumption.

Consider a few additional concepts before stating our main assumption. For each member of the two groups $g = A, B$, an outcome variable Y_{gi} and some individual characteristics X_i are observed. Y_g and X have a conditional joint distribution, $F_{Y_g,\mathbf{X}|D_g}(\cdot,\cdot) : \mathbb{R} \times \mathcal{X} \to [0, 1]$, and $\mathcal{X} \subset \mathbb{R}^k$ is the support of X.

The distribution of $Y_g|D_g$ is defined using the law of iterated probabilities, that is, after we integrate over the observed characteristics we obtain

$$F_{Y_g|D_g}(y) = \int F_{Y_g|X,D_g}(y|X = x) \cdot \mathrm{d}F_{X|D_g}(x), \quad g = A, B. \tag{4}$$

We can construct a counterfactual marginal wage distribution that mixes the conditional distribution of Y_A given X and $D_A = 1$ using the distribution of $X|D_B$. We denote that counterfactual distribution as $F_{Y_A^C:X=X|D_B}$, which is the distribution of

wages that would prevail for group B workers if they were paid like group A workers. This counterfactual distribution is obtained by replacing $F_{Y_B|X,D_B}$ with $F_{Y_A|X,D_A}$ (or $F_{X|D_A}$ with $F_{X|D_B}$) in Eq. (4):

$$F_{Y_A^C:X=X|D_B} = \int F_{Y_A|X,D_A}\left(y|X=x\right) \cdot \mathrm{d}F_{X|D_B}\left(x\right). \tag{5}$$

These types of manipulations play a very important role in the implementation of decomposition methods. Counterfactual decomposition methods can either rely on manipulations of F_X, as in DiNardo et al. (1996), or of $F_{Y|X}$, as in Albrecht et al. (2003) and Chernozhukov et al. (2009).[9]

Back to our union example, $F_{Y_B|X,D_B}\left(y|X=x\right)$ represents the conditional distribution of wages observed in the union sector, while $F_{Y_A|X,D_A}\left(y|X=x\right)$ represents the conditional distribution of wages observed in the non-union sector. In the case where $g=B$, Eq. (4) yields, by definition, the wage distribution in the union sector where we integrate the conditional distribution of wages given X over the marginal distribution of X in the union sector, $F_{X|D_B}\left(x\right)$. The counterfactual wage distribution $F_{Y_A^C:X=X|D_B}$ is obtained by integrating over the conditional distribution of wages in the non-union sector instead (Eq. (5)). It represents the distribution of wages that would prevail if union workers were paid like non-union workers.

The connection between these conditional distributions and the wage structure is easier to see when we rewrite the distribution of wages for each group in terms of the corresponding structural forms,

$$F_{Y_g|X,D_g}\left(y|X=x\right) = \Pr\left(m_g\left(X,\varepsilon\right) \leq y|X=x, D_g=1\right), \quad g=A,B.$$

Conditional on X, the distribution of wages only depends, therefore, on the conditional distribution of ε, and the wage structure $m_g\left(\cdot\right)$.[10] When we replace the conditional distribution in the union sector, $F_{Y_B|X,D_B}\left(y|X=x\right)$, with the conditional distribution in the non-union sector, $F_{Y_A|X,D_B}\left(y|X=x\right)$, we are replacing both the wage structure and the conditional distribution of ε. Unless we impose some further assumptions on the conditional distribution of ε, this type of counterfactual exercise will not yield interpretable results as it will mix differences in the wage structure and in the distribution of ε.

[9] Chernozhukov et al. (2009) discuss the conditions under which the two types of decomposition are equivalent.

[10] To see more explicitly how the conditional distribution $F_{Y_g|X,D_g}\left(\cdot\right)$ depends on the distribution of ε, note that we can write $F_{Y_g|X,D_g}\left(y|X=x\right) = \Pr\left(\varepsilon \leq m_g^{-1}\left(X,y\right)|X=x, D_g=1\right)$ under the assumption that $m(\cdot)$ is monotonic in ε (see Assumption 9 introduced below).

To see this formally, note that unless ε has the *same conditional distribution across groups*, the difference

$$\begin{aligned} F_{Y_B|D_B} - F_{Y_A^C:X=X|D_B} &= \int (\Pr(Y \leq y|X=x, D_B=1) \\ &\quad - \Pr(Y \leq y|X=x, D_A=1)) \cdot \mathrm{d}F_{X|D_B}(x) \\ &= \int (\Pr(m_B(X,\varepsilon) \leq y|X=x, D_B=1) \\ &\quad - \Pr(m_A(X,\varepsilon) \leq y|X=x, D_A=1)) \cdot \mathrm{d}F_{X|D_B}(x) \quad (6) \end{aligned}$$

will mix differences in m functions and differences in the conditional distributions of ε given X.

We are ultimately interested in a functional ν (i.e. a distributional statistic) of the wage distribution. The above result means that, in general, $\Delta_S^\nu \neq \nu(F_{Y_B|D_B}) - \nu(F_{Y_A^C:X=X|D_B})$. The question is under what additional assumptions will the difference between a statistic from the original distribution of wages and the counterfactual distribution, $\Delta_S^\nu = \nu(F_{Y_B|D_B}) - \nu(F_{Y_A^C:X=X|D_B})$, solely depend on differences in the wage structure? The answer is that under a conditional independence assumption, also known as *ignorability of the treatment* in the treatment effects literature, we can identify Δ_S^ν and the remaining terms Δ_X^ν and Δ_ε^ν.

Assumption 5 (*Conditional Independence/Ignorability*). For $g = A, B$, let (D_g, X, ε) have a joint distribution. For all x in $\mathcal{X}$: ε is independent of D_g given $X = x$ or, equivalently, $D_g \perp\!\!\!\perp \varepsilon|X$.

In the case of the simple counterfactual treatment, the identification restrictions from the treatment effects literature may allow the researcher to give a causal interpretation to the results of the decomposition methodology as discussed in Section 2.3. The ignorability assumption has become popular in empirical research following a series of papers by Rubin and coauthors and by Heckman and coauthors.[11] In the program evaluation literature, this assumption is sometimes called *unconfoundedness* or *selection on observables*, and allows identification of the treatment effect parameter.

2.1.5. Identification of the aggregate decomposition

We can now state our main result regarding the identification of the aggregate decomposition

Proposition 1 (***Identification of the Aggregate Decomposition***). *Under Assumption 3 (simple counterfactual), 4 (overlapping support), and 5 (ignorability), the overall ν-gap, Δ_O^ν, can*

[11] See, for instance, Rosenbaum and Rubin (1983, 1984), Heckman et al. (1997a,b) and Heckman et al. (1998).

be written as

$$\Delta_O^{\nu} = \Delta_S^{\nu} + \Delta_X^{\nu},$$

where

(i) *the wage structure term* $\Delta_S^{\nu} = \nu(F_{Y_B|D_B}) - \nu(F_{Y_A^C:X=X|D_B})$ *solely reflects the difference between the structural functions* $m_B(\cdot,\cdot)$ *and* $m_A(\cdot,\cdot)$

(ii) *the composition effect term* $\Delta_X^{\nu} = \nu(F_{Y_A^C:X=X|D_B}) - \nu(F_{Y_A|D_A})$ *solely reflects the effect of differences in the distribution of characteristics (*X *and* ε*) between the two groups.*

This important result means that, under the ignorability and overlapping assumptions, we can give a structural interpretation to the aggregate decomposition that is formally linked to the underlying wage setting models, $Y_A = m_A(X,\varepsilon)$ and $Y_B = m_B(X,\varepsilon)$. Note also that the wage structure (Δ_S^{ν}) and composition effect (Δ_X^{ν}) terms represent algebraically what we have informally defined by terms D.1⋆ and D.3.

As can be seen from Eq. (6), the only source of difference between $F_{Y_B|D_B}$ and $F_{Y_A^C:X=X|D_B}$ is the difference between the structural functions $m_B(\cdot)$ and $m_A(\cdot)$. Now note that under Assumptions 4 and 5, we have that $\Delta_O^{\nu} = \Delta_S^{\nu} + \nu(F_{Y_A^C:X=X|D_B}) - \nu(F_{Y_A|D_A})$, where

$$\begin{aligned} &F_{Y_A^C:X=X|D_B} - F_{Y_A|D_A} \\ &\quad = \int \Pr(Y \le y|X=x, D_A=1) \cdot \left(\mathrm{d}F_{X|D_B}(x) - \mathrm{d}F_{X|D_A}(x)\right). \end{aligned}$$

Thus, $\nu(F_{Y_A^C:X=X|D_B}) - \nu(F_{Y_A|D_A})$ reflects only changes or differences in the distribution of observed covariates. As a result, under Assumptions 4 and 5, we identify Δ_X^{ν} by $\nu(F_{Y_A^C:X=X|D_B}) - \nu(F_{Y_A|D_A})$ and set $\Delta_{\varepsilon}^{\nu} = 0$. This normalization makes sense as a result of the conditional independence assumption: no difference in wages will be systematically attributed to differences in distributions of ε once we fix these distributions to be the same given X. Thus, all remaining differences beyond Δ_S^{ν} are due to differences in the distribution of covariates captured by Δ_X^{ν}.

Combining these two results, we get

$$\begin{aligned} \Delta_O^{\nu} &= \left[\nu(F_{Y_B|D_B}) - \nu(F_{Y_A^C:X=X|D_B})\right] + \left[\nu(F_{Y_A^C:X=X|D_B}) - \nu(F_{Y_A|D_A})\right] \\ &= \Delta_S^{\nu} + \Delta_X^{\nu} \end{aligned} \tag{7}$$

which is the main result in Proposition 1.

When the Assumption 3 (simple counterfactual) and 5 (ignorability) are satisfied, the conditional distribution of Y given X remains invariant under manipulations of the

marginal distribution of X. It follows that Eq. (5) represents a valid counterfactual for the distribution of Y that would prevail if workers in group B were paid according to the wage structure $m_A(\cdot)$. The intuition for this result is simple. Since $Y_A = m_A(X, \varepsilon)$, manipulations of the distribution of X can only affect the conditional distribution of Y_A given X if they either (i) change the wage setting function $m_A(\cdot)$, or (ii) change the distribution of ε given X. The first change is ruled out by the assumption of a simple counterfactual treatment (i.e. no general equilibrium effects), while the second effect is ruled out by the ignorability assumption.

In the inequality literature, the invariance of the conditional distribution is often introduced as the key assumption required for $F_{Y_A^C:X=X|D_B}$ to represent a valid counterfactual (e.g. DiNardo et al., 1996; Chernozhukov et al., 2009).

Assumption 6 (*Invariance of Conditional Distributions*). The construction of the counterfactual wage distribution for workers of group B that would have prevailed if they were paid like group A workers (described in Eq. (5)), assumes that the conditional wage distribution $F_{Y_A|X,D_A}(y|X = x)$ applies or can be extrapolated for $x \in \mathcal{X}$, that is, it remains valid when the marginal distribution $F_{X|D_B}$ replaces $F_{X|D_A}$.

One useful contribution of this chapter is to show the economics underneath this assumption, i.e. that the invariance assumption holds provided that there are no general equilibrium effects (ruled out by Assumption 3) and no selection based on unobservables (ruled out by Assumption 5).

Assumption 6 is also invoked by Chernozhukov et al. (2009) to perform the aggregate decomposition using the following alternative counterfactual that uses group B as the reference group. Let $F_{Y_B^C:X=X|D_A}$ be the distribution of wages that would prevail for group A workers under the conditional distribution of wages of group B workers. In our union example, this would represent the distribution of wages of non-union workers that would prevail if they were paid like union workers. Relative to Eq. (7), the terms of the decomposition equation are now inverted:

$$\begin{aligned}\Delta_O^v &= \left[\nu(F_{Y_B|D_B}) - \nu(F_{Y_B^C:X=X|D_A})\right] + \left[\nu(F_{Y_B^C:X=X|D_A}) - \nu(F_{Y_A|D_A})\right] \\ &= \Delta_X^v + \Delta_S^v.\end{aligned}$$

Now the first term Δ_X^v is the composition effect and the second term Δ_S^v the wage structure effect.

Whether the assumption of the invariance of the conditional distribution is likely to be satisfied in practice depends on the economic context. If group A were workers in 2005 and group B were workers in 2007, perhaps Assumption 6 would be more likely to hold than if group A were workers in 2007 and group B were workers in 2009 in the presence of the 2009 recession. Thus it is important to provide an economic rationale

to justify Assumption 6 in the same way the choice of instruments has to be justified in terms of the economic context when using an instrumental variable strategy.

2.1.6. Why ignorability may not hold, and what to do about it

The conditional independence assumption is a somewhat strong assumption. We discuss three important cases under which it may not hold:

1. *Differential selection into labor market*. This is the selection problem that Heckman (1979) is concerned with in describing the wage offers for women. In the case of the gender pay gap analysis, it is quite plausible that the decisions to participate in the labor market are quite different for men and women. Therefore, the conditional distribution of $(X, \varepsilon)\,|D_B = 1$ may be different from the distribution of $(X, \varepsilon)\,|D_B = 0$. In that case, both the observed and unobserved components may be different, reflecting the fact that men participating in the labor market may be different in observable and unobservable ways from women who also participate. The ignorability assumption does not necessarily rule out the possibility that these distributions are different, but it constrains their relationship. Ignorability implies that the joint densities of observables and unobservables for groups A and B (men and women) have to be similar up to a ratio of conditional probabilities:

$$\begin{aligned} f_{X,\varepsilon|D_B}(x, e|1) &= f_{X,\varepsilon|D_B}(x, e|0) \cdot f_{X|D_B}(x|1)\,/f_{X|D_B}(x|0) \\ &= f_{X,\varepsilon|D_B}(x, e|0) \cdot \left(\frac{\Pr(D_B = 1|X = x)}{\Pr(D_B = 0|X = x)}\right) \cdot \left(\frac{\Pr(D_B = 0)}{\Pr(D_B = 1)}\right). \end{aligned}$$

2. *Self-selection into groups A and B based on unobservables*. In the gender gap example there is no selection into groups, although the consequences of differential selection into the labor market are indeed the same. An example where self-selection based on unobservables may occur is in the analysis of the union wage gap. The conditional independence or ignorability assumption rules out selection into groups based on unobservable components ε beyond X. However, the ignorability assumption does not impose that $(X, \varepsilon) \perp\!\!\!\perp D_B$, so the groups may have different marginal distributions of ε. But if selection into groups is based on unobservables, then the ratio of conditional joint densities will in general depend on the value of e being evaluated, and not only on x, as ignorability requires:

$$\frac{f_{X,\varepsilon|D_B}(x, e|1)}{f_{X,\varepsilon|D_B}(x, e|0)} \neq \left(\frac{\Pr(D_B = 1|X = x)}{\Pr(D_B = 0|X = x)}\right) \cdot \left(\frac{\Pr(D_B = 0)}{\Pr(D_B = 1)}\right).$$

3. *Choice of X and ε*. In the previous case, the values of X and ε are not determined by group choice, although they will be correlated and may even explain the choice of the group. In the first example of the gender pay gap, values of X and ε such as occupation choice and unobserved effort may also be functions of gender 'discrimination'. Thus,

the conditional independence assumption will not be valid if ε is a function of D_g, even holding X constant. The interpretation of ignorability here is that given the choice of X, the choice of ε will be randomly determined across groups. Pursuing the gender pay gap example, fixing X (for example education), men and women would exert the same level of effort. The only impact of anticipated discrimination is that they may invest differently in education.

In Section 6, we discuss several solutions to these problems that have been proposed in the decomposition literature. Those include the use of panel data methods or standard selection models. In case 2 above, one could also use instrumental variable methods to deal with the fact that the choice of group is endogenous. One identification issue we briefly address here is that IV methods would indeed yield a valid decomposition, but only for the subpopulation of compliers.

To see this, consider the case where we have a binary instrumental variable Z, which is independent of (ε, T) conditional on X, where T is a categorical variable which indicates 'type'. There are four possible types: a, n, c and d as described below:

Assumption 7 (*LATE*). For $g = A, B$, let (D_g, X, Z, ε) have a joint distribution in $\{0, 1\} \times \mathcal{X} \times \{0, 1\} \times \mathcal{E}$. We define T, a random variable that may take on four values $\{a, n, c, d\}$, and that can be constructed using D_B and Z according to the following rule: if $Z = 0$ and $D_B = 0$, then $T \in \{n, c\}$; if $Z = 0$ and $D_B = 1$, then $T \in \{a, d\}$; if $Z = 1$ and $D_B = 0$, then $T \in \{n, d\}$; if $Z = 1$ and $D_B = 1$, then $T \in \{a, c\}$.

(i) For all x in $\mathcal{X}$: Z is independent of (ε, T).

(ii) $\Pr(T = d|X = x) = 0$.

These are the LATE assumptions from Imbens and Angrist (1994), which allow us to identify the counterfactual distribution of $Y_A^C|X, D_B = 1, T = c$. We are then able to decompose the ν-wage gap under that less restrictive assumption, but only for the population of compliers:

$$\begin{aligned}\Delta^{\nu}_{O|T=c} &= \left[\nu(F_{Y_B|D_B,T=c}) - \nu(F_{Y_A^C:X=X|D_B,T=c})\right] \\ &\quad + \left[\nu(F_{Y_A^C:X=X|D_B,T=c}) - \nu(F_{Y_A|D_A,T=c})\right] \\ &= \Delta^{\nu}_{S|T=c} + \Delta^{\nu}_{X|T=c}.\end{aligned}$$

2.2. Case 2: The detailed decomposition

One convenient feature of the aggregate decomposition is that it can be performed without any assumption on the structural functional forms, $m_g(X, \varepsilon)$, while constraining the

distribution of unobserved (ε) characteristics.[12] Under the assumptions of Proposition 1, the composition effect component Δ_X^v reflects differences in the distribution of X, while the wage structure component Δ_S^v reflects differences in the returns to either X or ε.

To perform a detailed decomposition, we need to separate the respective contributions of X or ε in both Δ_S^v and Δ_X^v, in addition to separating the individual contribution of each element of the vector of covariates X. Thus, generally speaking, the identification of an interpretable detailed decomposition involves stronger assumptions such as functional form restrictions and/or further restrictions on the distribution of ε, like independence with respect to X and D.

Since these restrictions tend to be problem specific, it is not possible to present a general identification theory as in the case of the aggregate decomposition. We discuss instead how to identify the elements of the detailed decomposition in a number of specific cases. Before discussing these issues in detail, it is useful to state what we seek to recover with a detailed decomposition.

Property 1 (***Detailed Decomposition***). *A procedure is said to provide a detailed decomposition when it can apportion the composition effect, Δ_X^v, or the wage structure effect, Δ_S^v, into components attributable to each explanatory variable:*

1. *The contribution of each covariate X_k to the composition effect, $\Delta_{X_k}^v$, is the portion of Δ_X^v that is only due to differences between the distribution of X_k in groups A and B. When $\Delta_X^v = \sum_{k=1}^{K} \Delta_{X_k}^v$, the detailed decomposition of the composition effect is said to* ***add up****.*
2. *The contribution of each covariate X_k to the wage structure effect, $\Delta_{S_k}^v$, is the portion of Δ_S^v that is only due to differences in the parameters associated with X_k in group A and B, i.e. to differences in the parameters of $m_A(\cdot,\cdot)$ and $m_B(\cdot,\cdot)$ linked to X_k. Similarly, the contribution of unobservables ε to the wage structure effect, $\Delta_{S_\varepsilon}^v$, is the portion of Δ_S^v that is only due to differences in the parameters associated with ε in $m_A(\cdot,\cdot)$ and $m_B(\cdot,\cdot)$.*

Note that unobservables do not make any contribution to the composition effect because of the ignorability assumption we maintain throughout most of the chapter. As we mentioned earlier, it is also far from clear how to divide the parameters of the functions $m_A(\cdot,\cdot)$ and $m_B(\cdot,\cdot)$ into those linked to a given covariate or to unobservables. For instance, in a model with a rich set of interactions between observables and unobservables, it is not obvious which parameters should be associated with a given covariate. As a result, computing the elements of the detailed decomposition for the wage structure involves arbitrary choices to be made depending on the economic question of interest.

The adding-up property is automatically satisfied in linear settings like the standard OB decomposition, or the RIF-regression procedure introduced in Section 5.2. However, it is unlikely to hold in non-linear settings when the distribution of each individual

[12] Differences in the distribution of the ε are fairly constrained under the ignorability assumption. While the unconditional distribution of ε may differ between group A and B (because of differences in the distribution of X), the conditional distribution of ε has to be the same for the two groups.

covariate X_k is changed while keeping the distribution of the other covariates unchanged (e.g. in the case discussed in Section 5.3). In such a procedure "with replacement" we would, for instance, first replace the distribution of X_1 for group A with the distribution of X_1 for group B, then switch back to the distribution of X_1 for group A and replace the distribution of X_2 instead, etc.

By contrast, adding up would generally be satisfied in a sequential (e.g. "without replacement") procedure where we first replace the distribution of X_1 for group A with the distribution of X_1 for group B, and then do the same for each covariate until the whole distribution of X has been replaced. The problem with this procedure is that it would introduce some path dependence in the decomposition since the "effect" of changing the distribution of one covariate generally depends on distribution of the other covariates.

For example, the effect of changes in the unionization rate on inequality may depend on the industrial structure of the economy. If unions have a particularly large effect in the manufacturing sector, the estimated effect of the decline in unionization between, say, 1980 and 2000 will be larger under the distribution of industrial affiliation observed in 1980 than under the distribution observed in 2000. In other words, the order of the decomposition matters when we use a sequential (without replacement) procedure, which means that the property of path independence is violated. As we will show later in the chapter, the lack of path independence in many existing detailed decomposition procedures based on a sequential approach is an important shortcoming of these approaches.

Property 2 (***Path Independence***). *A decomposition procedure is said to be path independent when the order in which the different elements of the detailed decomposition are computed does not affect the results of the decomposition.*

A possible solution to the problem of path dependence suggested by Shorrocks (1999) consists of computing the marginal impact of each of the factors as they are eliminated in succession, and then averaging these marginal effects over all the possible elimination sequences. He calls the methodology the Shapley decomposition, because the resulting formula is formally identical to the Shapley value in cooperative game theory. We return to these issues later in the chapter.

2.2.1. Nonparametric identification of structural functions

One approach to the detailed decomposition is to identify the structural functions $m_A(\cdot,\cdot)$ and $m_B(\cdot,\cdot)$, and then use the knowledge of these structural forms to compute various counterfactuals of interest. For example, one could look at what happens when all the parameters of $m_A(\cdot,\cdot)$ pertaining to education are switched to their values estimated for group B, while the rest of the $m_A(\cdot,\cdot)$ function remains unchanged.

For the purpose of identifying the structural functions $m_A(\cdot,\cdot)$ and $m_B(\cdot,\cdot)$, neither ignorability nor LATE assumptions are very helpful. Stronger assumptions invoked in the

literature on nonparametric identification of structural functions (e.g. Matzkin, 2003; Blundell and Powell, 2007; Imbens and Newey, 2009) have to be used instead:

Assumption 8 (*Independence*). For $g = A, B$, $X \perp\!\!\!\perp \varepsilon | D_g$.

Assumption 9 (*Strict Monotonicity in the Random Scalar* ε). For $g = A, B$ and for all values x in $\mathcal{X}$, ε is a scalar random variable and $m_g(X, \varepsilon)$ is strictly increasing in ε.

With these two additional assumptions we can write, for $g = A, B$, the functions $m_g(\cdot, \cdot)$ using solely functionals of the joint distribution of (Y, D_g, X). We can assume without loss of generality that $\varepsilon | D_g \sim U[0, 1]$, because (i) we observe the conditional distributions of $X | D_g$, and ε is a scalar random variable independent of X given D_g. Once we have identified the functions $m_g(\cdot, \cdot)$ for $g = A, B$, we can construct the counterfactual distribution of $F_{Y_A^C : X = X | D_B}$ and compute any distributional statistic of interest.[13]

Note, however, that the monotonicity assumption is not innocuous in the context of comparisons across groups. If there was only one group of workers, the monotonicity assumption would be a simple normalization. With more than one group, however, it requires that the same unobservable variable has positive returns for all groups of workers, which in some settings may not be plausible, though this is automatically satisfied in additively separable models.

There are various reasons why this assumption may be problematic in practice. Empirical wage distributions exhibit many flat spots because of heaping or minimum wage effects. For example, if group A and B corresponded to two different years or countries with different minimum wages, the monotonicity assumption would not be satisfied.[14] The monotonicity assumption would also break down in the presence of measurement error in wages since the wage residual would now mix measurement error and unobserved skills. As a result, the same amount of unobserved skills would not guarantee the same position in the conditional distribution of residuals in the two groups.

In most labor economics applications, assuming that unobservables are independent of the covariates is a strong and unrealistic assumption. Thus, the identification of the structural functions comes at a relatively high price. The milder assumption of ignorability allows us to identify Δ_S^v and Δ_X^v. With full independence, we can go back and identify more terms. In fact, because we obtain an expression for Δ_S^v, we can construct detailed decompositions by fixing deterministically the values of some covariates while letting others vary.

[13] This monotonicity assumption can also be found in the works of Matzkin (2003), Altonji and Matzkin (2005), Imbens and Newey (2009), and Athey and Imbens (2006).

[14] The rank pairing of two outcome variables Y_A and Y_B will be disrupted if the rank of Y_A remains the same because at a mass point corresponding to the minimum wage, while the rank of Y_B continues to increase in the absence of minimum wage at the rank. Heckman et al. (1997a,b) consider the case of mass points at zero, but the case of multiple mass points is much more difficult.

2.2.2. Functional form restrictions: decomposition of the mean

A more common approach used in the decomposition literature consists of imposing functional form restrictions to identify the various elements of a detailed decomposition. For instance, detailed decompositions can be readily computed in the case of the mean using the assumptions implicit in Oaxaca (1973) and Blinder (1973). The first assumption is additive linearity of the $m_g(\cdot,\cdot)$ functions. The linearity assumption is also commonly used in quantile-based decomposition methodologies, such as Albrecht et al. (2003), Machado and Mata (2005), and Melly (2006). The linearity assumption allows for heteroscedasticity due, for example, to the fact that the variance of unobservables increases as educational attainment increases.

Assumption 10 (*Additive Linearity*). The wage structure, m_A and m_B, are linear additively separable functions in the worker's observable and unobservable characteristics:

$$Y_{gi} = m_g(X_i, \varepsilon_i) = X_i\beta_g + \upsilon_{ig}, \quad g = A, B$$

where $\upsilon_{ig} = h_g(\varepsilon_i)$.

The second assumption implicit in the OB procedure is that the conditional mean of υ_{ig} is equal to zero:

Assumption 11 (*Zero Conditional Mean*). $\mathbb{E}[\upsilon_g|X, D_B] = 0$.

Under mean independence, we have that for $g = A, B$, $\mathbb{E}[Y_g|D_g = 1] = \mathbb{E}[X|D_g = 1]\beta_g$ and therefore we can write the mean counterfactual $\mu(F_{Y_A^C:X=X|D_B})$ as $\mathbb{E}[X|D_B = 1]\beta_A$. Therefore,

$$\Delta_S^\mu = \mathbb{E}[X|D_B = 1](\beta_B - \beta_A) \quad \text{and} \quad \Delta_X^\mu = (\mathbb{E}[X|D_B = 1] - \mathbb{E}[X|D_B = 0])\beta_A.$$

2.2.3. Functional form restrictions: more general decompositions

Under Assumption 11, the error term conveniently drops out of the decomposition for the mean. For more general distributional statistics such as the variance, however, we need more assumptions about the distribution of unobservables to perform a detailed decomposition. If we add the following assumptions on the conditional wage variance and on the function of the unobservables υ_{ig}, we can separate out the wage structure effects of observables and unobservables.

Assumption 12 (*Constant Returns to Unobservables*). For $g = A, B$, $\upsilon_g = \sigma_g\varepsilon$.

Assumption 13 (*Homoscedasticity*). For $g = A, B$, $\text{Var}[\varepsilon|X, D_g = 1] = 1$.

Under these two additional assumptions, we can identify σ_g, and interpret it as the price of unobservables.[15] Assumption 10 (additive linearity) then allows us to separate

out returns to observable and unobservable factors, and to separately identify the contribution of observable and unobservable factors to the wage structure effect. Note that because of the zero conditional mean assumption, only the observable factors influence mean wages.

More formally, consider the counterfactual wage, $Y_A^{C,1}$, for group B workers where the return to unobservables is set to be as in group A,[16]

$$Y_A^{C,1} = X\beta_B + \sigma_A \varepsilon. \tag{8}$$

Under the Assumption 5, and 9 to 13, we can divide the wage structure effect into a component linked to unobservables, $\Delta_{S,\sigma}^{\nu}$, and a component linked to observables, $\Delta_{S,\beta}^{\nu}$, as follows

$$\Delta_S^{\nu} = \underbrace{\left[\nu(F_{Y_B|D_B}) - \nu(F_{Y_A^{C,1}:X=X|D_B})\right]}_{\Delta_{S,\sigma}^{\nu}} + \underbrace{\left[\nu(F_{Y_A{}^{C,1}:X=X|D_B}) - \nu(F_{Y_A{}^{C}:X=X|D_B})\right]}_{\Delta_{S,\beta}^{\nu}}.$$

The above assumptions correspond to those implicitly used by Juhn et al. (1991) in their influential study on the evolution of the black-white wage gap.[17] While it is useful to work with a single "price" of unobservables σ_g, doing so is not essential for performing a detailed decomposition. Juhn et al. (1993) [JMP] use a weaker set of assumptions in their influential study of wage differentials over time that we now discuss in more detail.

JMP propose a residual imputation procedure that relies on the key assumption that the rank of worker i in the distribution of υ_A is the same as in the distribution of υ_B, conditional on X. This procedure enables them to perform a decomposition even when the function $h_g(\cdot)$ used to define the regression error $\upsilon_g = h_g(\varepsilon)$ is not linear (non-linear skill pricing). Since the (conditional) rank of υ_g normalized on a [0, 1] scale is simply the cumulative distribution $F_{\upsilon_B|X}(\cdot)$ evaluated at that point, conditional rank preservation can be stated as follows in our context:

Assumption 14 (*Conditional Rank Preservation*). For all individual i, we have $\tau_{Ai}(x_i) = \tau_{Bi}(x_i)$, where $\tau_{Ai}(x_i) = F_{\upsilon_A|X}(\upsilon_{Ai}|X = x_i)$ and $\tau_{Bi}(x_i) = F_{\upsilon_B|X}(\upsilon_{Bi}|X = x_i)$ are the rankings of υ_{Ai} and υ_{Bi} in their respective conditional distributions.

15 Note that it is possible to relax the homoskedasticity assumption while maintaining the assumption of a single price of unobservables σ_g, as in Chay and Lee (2000). We do not follow this approach here to simplify the presentation.

16 Note that we depart somewhat from our previous notation, as $Y_A^{C,1}$ retains some components of the structural form of group B, which will disappear in $Y_A^{C,3}$ below.

17 See Blau and Kahn (1992, 2003) for an application of the methodology to the study of gender wage differentials across countries.

Under this assumption, if individual i in group A observed at rank $F_{\upsilon_A|X}(\upsilon_{iA}|X = x_i)$ were in group B instead, he/she would remain at the same rank in the conditional distribution of υ for that other group (and vice versa). Conditional rank preservation is a direct consequence of the assumptions of ignorability (Assumption 5) and monotonicity (Assumption 9). Under ignorability, the distribution of ε given X does not depend on group membership. Since $\upsilon_A = h_A(\varepsilon)$ and $\upsilon_B = h_B(\varepsilon)$, the rank of υ_A and υ_A in their respective distributions is the same as the rank of ε, provided that $h_A(\cdot)$ and $h_B(\cdot)$ are monotonic.

Note that the assumption of rank preservation is substantially stronger than ignorability. For instance, consider the case where ε is a vector of two ability measures: cognitive ability and manual ability. If cognitive ability is more valued under the wage structure $m_A(\cdot)$ than under the wage structure $m_B(\cdot)$, the ranking of workers in the A and B distributions will be different, which means that neither monotonicity nor rank preservation will hold. But provided that the conditional distribution of cognitive and manual ability given X is the same for groups A and B, ignorability holds, which means that the aggregate decomposition is still identified.

We explain how to implement the JMP procedure in practice in Section 4.3. Compared to the procedure described above to construct the counterfactual wage, $Y_A^{C,1} = X\beta_B + \sigma_A\varepsilon$, the difference is that an imputed residual from the group A distribution is used instead of $\sigma_A\varepsilon$. The idea is to replace υ_{Bi} with rank $\tau_{Bi}(x_i)$ in the conditional distribution of υ_B with an imputed error term

$$\upsilon_{Ai}^{C,2} = F^{-1}_{\upsilon_A|X}(\tau_{Bi}(x_i), x_i). \tag{9}$$

The resulting counterfactual wage for group B workers,

$$Y_{Ai}^{C,2} = X\beta_B + \upsilon_{Ai}^{C,2}, \tag{10}$$

can then be used to compute the following two elements of the decomposition:

$$\Delta^{\nu}_{S,\sigma} = \nu(F_{Y_B|D_B}) - \nu(F_{Y_A^{C,2}:X=X|D_B}) \quad \text{and}$$
$$\Delta^{\nu}_{S,\beta} = \nu(F_{Y_A^{C,2}:X=X|D_B}) - \nu(F_{Y_A^{C}:X=X|D_B}).$$

One important implementation issue we discuss in Section 4.3 is how to impute residuals conditional on X. This is an important limitation of JMP's procedure that can be addressed in a number of ways. One popular approach is to use conditional quantile regressions to allow for different returns to observables that vary along the conditional wage distribution. This approach was proposed by Machado and Mata (2005) and reexamined by Albrecht et al. (2003) and Melly (2005). It relies on the assumption that the

conditional distribution of $Y_g|X, D_g$, is completely characterized by the collection of regression quantiles $\{\beta_{g,\tau}; \tau \in (0,1)\}$.

Assumption 15 (*Heterogenous Returns to Observables*). For $g = A, B$, $Y_{gi} = X_i\beta_{g,\tau} + h_{g,\tau}(\varepsilon_i)$.

Assumption 16 (*Complete Collection of Linear Conditional Quantiles*). For $g = A, B$, and $\forall\tau \in (0,1)$ $\tau = \Pr(Y_g \leq x\beta_{g,\tau}|X = x, D_g = 1)$.

The above assumptions plus ignorability allow the decomposition of Δ_O^ν into Δ_S^ν and Δ_X^ν. Note that because $\tau = F_{Y_g|X,D_g}(x\beta_{g,\tau}|X = x)$ for all τ, we are fully parameterizing the conditional distribution of $Y_g|X, D_g$ by $\beta_{g,\tau}$ using all $\tau \in (0,1)$. Thus, once one inverts the conditional quantile to obtain a conditional CDF, one can apply Eq. (4) and (5) to compute an actual or counterfactual distribution.

Many other decomposition methods have been proposed to deal with parametric and nonparametric identification of conditional distribution functions. We have discussed the JMP procedure, as well as extensions to the case of conditional quantiles, as a way of illustrating the kind of assumptions required for identifying detailed decompositions of general distributional statistics. The general message is that more stringent assumptions have to be imposed to perform a detailed decomposition instead of an aggregate decomposition. The same general message would apply if we had discussed the identification of other decomposition procedures such as (to cite a few examples) Donald et al. (2000), Fortin and Lemieux (1998), Melly (2005), Chernozhukov et al. (2009), and Rothe (2009) instead.

Finally, it is also possible to relax some of the above assumptions provided that other assumptions are used instead. For instance, if one fixes the prices of unobservables to be the same across groups, say to a unit price, then $\Delta_{S,\sigma}^\nu$ reflects in fact changes in the distribution of unobservables. In that case, ignorability does not hold, but because of linearity and zero conditional mean assumptions we can identify the parameter β's. The difference between $\left(Q_{B,\tau}(X_i) - X_i\beta_B\right)$ and $\left(Q_{A,\tau}(X_i) - X_i\beta_A\right)$ is interpreted as differences in the τ-quantile of the conditional distribution of ε given X across groups B and A ($Q_{g,\tau}(X)$ is the τ-quantile of the conditional distribution of Y for group g). Let us state the following normalization assumption.

Assumption 17 (*Unit Price to Unobservables*). For $g = A, B$, $\upsilon_g = \sigma_g\varepsilon = \varepsilon$.

The overall wage gap can then be decomposed as follows

$$\begin{aligned}\Delta_O^\nu &= \Delta_S^\nu + \Delta_\varepsilon^\nu + \Delta_X^\nu \\ &= \underbrace{\left[\nu(F_{Y_B|D_B}) - \nu(F_{Y_A^C:(X,\varepsilon)=(X,\varepsilon)|D_B})\right]}_{\Delta_S^\nu}\end{aligned}$$

$$+ \underbrace{\left[\nu(F_{Y_A^C:(X,\varepsilon)=(X,\varepsilon)|D_B}) - \nu(F_{Y_A^C:X=X|D_B})\right]}_{\Delta_\varepsilon^\nu}$$
$$+ \underbrace{\left[\nu(F_{Y_A^C:X=X|D_B}) - \nu(F_{Y_A|D_A})\right]}_{\Delta_X^\nu}. \tag{11}$$

Because of Assumptions 10, 12 and 17, we now have $Y_A = X\beta_A + \varepsilon$ and $Y_B = X\beta_B + \varepsilon$. The first difference Δ_S^ν, corresponds to differences in β's only; the second difference is due to differences in

$$F_{Y_A^C:(X,\varepsilon)=(X,\varepsilon)|D_B} - F_{Y_A^C:X=X|D_B},$$

which are explained by differences in the conditional distribution of ε given X across groups B and A. Thus, an easy way to obtain that difference is to construct a counterfactual

$$Y_{Ai}^{C,3} = X_i\beta_A + (Y_{Bi} - X_i\beta_B), \tag{12}$$

and to replace $F_{Y_A^C:(X,\varepsilon)=(X,\varepsilon)|D_B}$ with $F_{Y_A^{C,3}:(X,\varepsilon)=(X,\varepsilon)|D_B}$ given that they will be equivalent under the above functional form assumptions.

Finally, the difference Δ_X^ν can be obtained as a residual difference. However, under the maintained assumptions it shall reflect only differences in the marginal distributions of X.

2.3. Decomposition terms and their relation to causality and the treatment effects literature

We end this section by discussing more explicitly the connection between decompositions and various concepts introduced in the treatment effects literature. As it turns out, when the counterfactuals are based on hypothetical alternative wage structures, they can be easily linked to the treatment effects literature. For example: What if group A workers were paid according to the wage structure of group B? What if all workers were paid according to the wage structure of group A?

Define the overall average treatment effect (*ATE*) as the difference between average wages if everybody were paid according to the wage structure of group B and average wages if everybody were paid according to the wage structure of group A. That is:

$$ATE = \mathbb{E}[Y_B] - \mathbb{E}[Y_A],$$

where switching a worker of from "type A" to "type B" is thought to be the "treatment".

We also define the average treatment effect on the treated (*ATT*) as the difference between actual average wages of group B workers and average wages if group B workers were paid according to the pay structure of group A. That is:

$$ATT = \mathbb{E}\left[Y_B | D_B = 1\right] - \mathbb{E}\left[Y_A | D_B = 1\right].$$

These treatment effects can be generalized to other functionals or statistics of the wage distribution. For example, define ν-*TE*, the ν-treatment effect, as

$$\nu\text{-}TE = \nu\left(F_{Y_B}\right) - \nu\left(F_{Y_A}\right),$$

and its version applied to the subpopulation of "treated", ν-*TT* as

$$\nu\text{-}TT = \nu\left(F_{Y_B|D_B}\right) - \nu\left(F_{Y_A|D_B}\right).$$

The distributions F_{Y_B}, F_{Y_A} and $F_{Y_A|D_B}$ are not observed from data on (Y, D_B, X).[18] Following the treatment effects literature, we could in principle identify these parameters if "treatment" was randomly assigned. This is hardly the case, at least for our examples, and one needs extra identifying restrictions. In fact, we note that ignorability and common support assumptions (which together are termed strong ignorability after (Rosenbaum and Rubin, 1983)) are sufficient to guarantee identification of the previous parameters. For example under strong ignorability, for $g = A, B$

$$\begin{aligned} F_{Y_g}(y) &= \mathbb{E}\left[F_{Y|X,D_g}(y|X)\right], \\ F_{Y_A|D_B}(y) &= \mathbb{E}\left[F_{Y|X,D_A}(y|X)\,|D_B = 1\right]. \end{aligned}$$

Under ignorability, it follows that $F_{Y_A|D_B} \sim F_{Y_A^C:X=X|D_B}$. Then $\Delta_S^\nu = \nu\text{-}TT$ and $\Delta_X^\nu = \nu(F_{Y_B|D_B}) - \nu(F_{Y_A|D_A}) - (\nu\text{-}TT)$. Reweighting methods, as discussed by DiNardo et al. (1996), Hirano et al. (2003) and Firpo (2007, 2010) have implicitly or explicitly assumed strong ignorability to identify specific ν-treatment effects.

It is interesting to see how the choice of the reference or base group is related to the treatment effects literature. Consider the treatment effect parameter for the non-treated, ν-*TNT*:

$$\nu\text{-}TNT = \nu\left(F_{Y_B|D_A}\right) - \nu\left(F_{Y_A|D_A}\right).$$

[18] Only $F_{Y_B|D_B}$ and $F_{Y_A|D_A}$ are observed.

Under strong ignorability, we have $F_{Y_B|D_A}(\cdot) = \mathbb{E}[F_{Y|X,D_B}(\cdot|X)|D_B = 0] = F_{Y_B^C:X=X|D_A}(\cdot)$. Thus, in this case, $\Delta_S^\nu = \nu\text{-}TNT$ and $\Delta_X^\nu = \nu(F_{Y_B|D_B}) - \nu(F_{Y_A|D_A}) - (\nu\text{-}TNT)$.

We could also consider other decompositions, such as:

$$\nu\left(F_{Y_B|D_B}\right) - \nu\left(F_{Y_A|D_A}\right) = \nu\text{-}TE + \left(\nu\left(F_{Y_B|D_B}\right) - \nu\left(F_{Y_B}\right)\right) + \left(\nu\left(F_{Y_A}\right) - \nu\left(F_{Y_A|D_A}\right)\right),$$

where F_{Y_B} includes the actual wages of group B workers and the counterfactual wages of group A workers if they were are paid like group B workers, and conversely for F_{Y_A}. In this case, the wage structure effect is $\nu\text{-}TE$, while the composition effect is the sum $(\nu(F_{Y_B|D_B}) - \nu(F_{Y_B})) + (\nu(F_{Y_A}) - \nu(F_{Y_A|D_A}))$.[19]

The above discussion reveals that the reference group choice problem is just a matter of choosing a meaningful counterfactual. There will be no right answer. In fact, we see that analogously to the treatment effects literature, where treatment effect parameters are different from each other because they are defined over distinct subpopulations, the many possible ways of performing decompositions will reflect the reference group that we want to emphasize.

We conclude this section by discussing briefly the relationship between causality, structural parameters and decomposition terms. In this section, we show that the decomposition terms do not necessarily rely on the identification of structural forms. Whenever we can identify those structural functions linking observable and unobservable characteristics to wages, we benefit from being able to perform counterfactual analysis that we may not be able to do otherwise. However, that comes at the cost of having to impose either strong independence assumptions, as in the case of nonparametric identification, or restrictive functional form assumptions plus some milder independence assumption (mean independence, for instance) between observables and unobservables within each group of workers.

If we are, however, interested in the aggregate decomposition terms Δ_X^ν and Δ_S^ν, we saw that a less restrictive assumption is sufficient to guarantee identification of these terms. Ignorability is the key assumption here as it allows fixing the conditional distribution of unobservables to be the same across groups. The drawback is that we cannot separate out the wage structure effects associated with particular observable and unobservable characteristics.

The treatment effects literature is mainly concerned with causality. Under what conditions can we claim that although identifiable under ignorability, Δ_S^ν may have a causal interpretation? The conditions under which we could say that Δ_S^ν is a causal

[19] We note that this last decomposition corresponds, in the OB context, to the so-called three-fold decomposition presented in footnote 3.

parameter are very stringent and unlikely to be satisfied in general cases. There are two main reasons for that, in our view.

First, in many cases, "treatment" is not a choice or a manipulable action. When decomposing gender or race gaps in particular, we cannot conceive workers choosing which group to belong to.[20] They may have different labor market participation behavior, which is one case where ignorability may not hold, as discussed in Section 2.1.6. However, workers cannot choose treatment. Thus, if we follow, for example, Holland (1986)'s discussion of causality, we cannot claim that Δ_S^v is a causal parameter.

A second reason for failing to assign causality to the pay structure effect is that most of the observable variables considered as our X (or unobservables ε) are not necessarily pre-treatment variables.[21] In fact, X may assume different values as a consequence of the treatment. In the treatment effects literature, a confounding variable X may have different distributions across treatment groups. But that is not a direct action of the treatment. It should only be a selection problem: People who choose to be in a group may have a different distribution of X relative to people who choose to be in the other group. When X is affected by treatment, we cannot say that controlling for X we will obtain a causal parameter. In fact, what we will obtain is a partial effect parameter, netted from the indirect effect through changes in X.

3. OAXACA-BLINDER—DECOMPOSITIONS OF MEAN WAGES DIFFERENTIALS

In this section, we review the basics of OB decompositions, discussing at length some thorny issues related to the detailed decomposition. We also address alternative choices of counterfactuals, including the case of the pooled regression that uses a group membership dummy to obtain a measure of the aggregate wage structure effect. We introduce a reweighted-regression decomposition as an attractive alternative when the linearity of the conditional mean as a function of the covariates is questionable. Finally, we briefly discuss the extensions of OB decompositions to limited dependent variable models, which carry some of the issues, such as path dependence, that will surface in methods that go beyond the mean.

3.1. Basics

Despite its apparent simplicity, there are many important issues of estimation and interpretation in the classic OB decomposition. The goal of the method is to decompose differences in mean wages, μ, across two groups. The wage setting model is assumed to be linear and separable in observable and unobservable characteristics (Assumption 10):

$$Y_g = X\beta_g + \upsilon_g, \quad \text{for } g = A, B \tag{13}$$

[20] The union/non-union wage gaps or private/public sector wage gaps are more amenable to choice.

[21] Note that some analyses (e.g. Neal and Johnson, 1996) take great care to focus on pre-market variables.

where $\mathbb{E}[\upsilon_g|X] = 0$ (Assumption 11). Letting $D_B = 1$ be an indicator of group B membership, and taking the expectations over X, the overall mean wage gap Δ_O^μ can be written as

$$\begin{aligned}\Delta_O^\mu &= \mathbb{E}[Y_B|D_B = 1] - \mathbb{E}[Y_A|D_B = 0] \\ &= \mathbb{E}[\mathbb{E}(Y_B|X, D_B = 1)|D_B = 1] - \mathbb{E}[\mathbb{E}(Y_A|X, D_B = 0)|D_B = 0] \\ &= (\mathbb{E}[X|D_B = 1]\beta_B + \mathbb{E}[\upsilon_B|D_B = 1]) - (\mathbb{E}[X|D_B = 0]\beta_A + \mathbb{E}[\upsilon_A|D_B = 0])\end{aligned}$$

where $\mathbb{E}[\upsilon_A|D_B = 0] = \mathbb{E}[\upsilon_B|D_B = 1] = 0$. Adding and subtracting the average counterfactual wage that group B workers would have earned under the wage structure of group A, $\mathbb{E}[X|D_B = 1]\beta_A$, the expression becomes

$$\begin{aligned}\Delta_O^\mu &= \mathbb{E}[X|D_B = 1]\beta_B - \mathbb{E}[X|D_B = 1]\beta_A \\ &\quad + \mathbb{E}[X|D_B = 1]\beta_A - \mathbb{E}[X|D_B = 0]\beta_A \\ &= \mathbb{E}[X|D_B = 1](\beta_B - \beta_A) + (\mathbb{E}[X|D_B = 1] - \mathbb{E}[X|D_B = 0])\beta_A \\ &= \Delta_S^\mu + \Delta_X^\mu.\end{aligned}$$

Replacing the expected value of the covariates $\mathbb{E}[X|D_B = d]$, for $d = 0, 1$, by the sample averages $\overline{X}_g$, the decomposition is estimated as

$$\widehat{\Delta}_O^\mu = \overline{X}_B\widehat{\beta}_B - \overline{X}_B\widehat{\beta}_A + \overline{X}_B\widehat{\beta}_A - \overline{X}_A\widehat{\beta}_A \tag{14}$$

$$= \overline{X}_B\left(\widehat{\beta}_B - \widehat{\beta}_A\right) + \left(\overline{X}_B - \overline{X}_A\right)\widehat{\beta}_A \tag{15}$$

$$= \widehat{\Delta}_S^\mu + \widehat{\Delta}_X^\mu. \tag{16}$$

The first term in Eq. (15) is the wage structure effect, $\widehat{\Delta}_S^\mu$, while the second term is the composition effect, $\widehat{\Delta}_X^\mu$. Note that in cases where group membership is linked to some immutable characteristics of the workers, such as race or gender, the wage structure effect has also been called the "unexplained" part of the wage differentials or the part due to "discrimination".

The OB decomposition is very easy to use in practice. It is computed by plugging in the sample means and the OLS estimates $\widehat{\beta}_g$ in the above formula. Various good implementations of the procedure are available in existing software packages.[22] Table 2 displays the various underlying elements of the decomposition in the case of the gender wage gap featured in O'Neill and O'Neill (2006) using data from the NLSY79. The composition effect is computed as the difference between the male and female means reported in column (1) multiplied by the male coefficients reported in column (2).[23]

[22] The empirical applications of the OB procedure in this chapter use Jann (2008) procedures in Stata.

Table 2 Means and OLS regression coefficients of selected variables from NLSY log wage regressions for workers ages 35-43 in 2000.

Explanatory variables	(1) Means		(2) Male coef.		(3) Female coef.		(4) Male coef.		(5) Pooled coef.	
Female	0	1							−0.092	(0.014)
Education and skill level										
<10 yrs.	0.053	0.032	−0.027	(0.043)	−0.089	(0.050)	−0.027	(0.043)	−0.045	(0.033)
10-12 yrs (no diploma or GED)	0.124	0.104								
HS grad (diploma)	0.326	0.298	−0.013	(0.028)	−0.002	(0.029)	−0.013	(0.028)	−0.003	(0.020)
HS grad (GED)	0.056	0.045	0.032	(0.042)	−0.012	(0.044)	0.032	(0.042)	0.006	(0.030)
Some college	0.231	0.307	0.164	(0.031)	0.101	(0.030)	0.164	(0.031)	0.131	(0.022)
BA or equiv. degree	0.155	0.153	0.380	(0.037)	0.282	(0.036)	0.380	(0.037)	0.330	(0.026)
MA or equiv. degree	0.041	0.054	0.575	(0.052)	0.399	(0.046)	0.575	(0.052)	0.468	(0.034)
Ph.D or prof. Degree	0.015	0.007	0.862	(0.077)	0.763	(0.100)	0.862	(0.077)	0.807	(0.060)
AFQT percentile score (x.10)	4.231	3.971	0.042	(0.004)	0.041	(0.004)	0.042	(0.004)	0.042	(0.003)
L.F. withdrawal due to family resp.	0.129	0.547	−0.078	(0.025)	−0.083	(0.019)	−0.078	(0.025)	−0.067	(0.015)
Lifetime work experience										
Years worked civilian	17.160	15.559	0.038	(0.003)	0.030	(0.002)	0.038	(0.003)	0.033	(0.002)
Years worked military	0.578	0.060	0.024	(0.005)	0.042	(0.013)	0.024	(0.005)	0.021	(0.004)
% worked part-time	0.049	0.135	−0.749	(0.099)	−0.197	(0.049)	−0.749	(0.099)	−0.346	(0.044)

(*continued on next page*)

Table 2 (continued)

Explanatory variables	(1) Means		(2) Male coef.		(3) Female coef.		(4) Male coef.		(5) Pooled coef.	
Industrial Sectors										
Primary, Constr. & Utilities	0.186	0.087					0.059	(0.031)		
Manufacturing	0.237	0.120	0.034	(0.026)	0.140	(0.035)	0.093	(0.029)	0.072	(0.021)
Education, Health, & Public Adm.	0.130	0.358	−0.059	(0.031)	0.065	(0.030)			−0.001	(0.020)
Other Services	0.447	0.436	0.007	(0.024)	0.088	(0.029)	0.066	(0.026)	0.036	(0.018)
Constant			2.993	(0.156)	2.865	(0.144)	2.934	(0.157)	2.949	(0.105)
Dependent Var. (Log Hourly Wage)	2.763	2.529								
Adj. R-Square			0.422		0.407		0.422		0.431	
Sample size	2655	2654								

The data is an extract from the NLSY79 used in O'Neill and O'Neill (2006). Industrial sectors were added (at a lost of 89 observations) to their analysis to illustrate issues linked to categorical variables. The other explanatory variables are age, dummies for black, hispanic, region, msa, central city. Standard errors are in parentheses.

The corresponding wage structure effect is computed from the difference between the male and female coefficients reported in columns (2) and (3). The results are reported in column (1) of Table 3. The composition effect accounts for 0.197 (0.018) log points out of the 0.233 (0.015) average log wage gap between men and women in 2000. When the male wage structure is used as reference, only an insignificant 0.036 (0.019) part of the gap (the wage structure effect) is left unexplained.

Because of the additive linearity assumption, it is easy to compute the various elements of the detailed decomposition. The wage structure and composition effects can be written in terms of sums over the explanatory variables

$$\widehat{\Delta}_S^{\mu} = (\widehat{\beta}_{B0} - \widehat{\beta}_{A0}) + \sum_{k=1}^{M} \overline{X}_{Bk}(\widehat{\beta}_{Bk} - \widehat{\beta}_{Ak}), \tag{17}$$

$$\widehat{\Delta}_X^{\mu} = \sum_{k=1}^{M} \left(\overline{X}_{Bk} - \overline{X}_{Ak}\right) \widehat{\beta}_{Ak}, \tag{18}$$

where $(\widehat{\beta}_{B0} - \widehat{\beta}_{A0})$ represents the omitted group effect, and where $\overline{X}_{gk}$ and $\widehat{\beta}_{gk}$ represent the kth element of $\overline{X}_g$ and $\widehat{\beta}_g$, respectively. $\left(\overline{X}_{Bk} - \overline{X}_{Ak}\right) \widehat{\beta}_{Ak}$ and $\overline{X}_{Bk}(\widehat{\beta}_{Bk} - \widehat{\beta}_{Ak})$ are the respective contributions of the kth covariate to the composition and wage structure effect. Each element of the sum $\widehat{\Delta}_S^{\mu}$ can be interpreted as the contribution of the difference in the returns to the kth covariate to the total wage structure effect, evaluated at the mean value of X_{Bk}. Whether or not this decomposition term is economically meaningful depends on the choice of the omitted group, an issue we discuss in detail in Section 3.2 below.[24]

Similar to O'Neill and O'Neill (2006), Table 3 reports the contribution of single variables and groups of variables to composition (upper panel) and wage structure effects (lower panel). Life-time work experience 'priced' at the male returns to experience stands out as the factor with the most explanatory power (0.137 out of 0.197, or 69%) for composition effects. The wage structure effects are not significant in this example, except for the case of industrial sectors which we discuss below.

Because regression coefficients are based on partial correlations, an OB decomposition that includes all K explanatory variables of interest satisfies the property of path independence (Property 2). Note, though, that a sequence of Oaxaca-Blinder decompositions, each including a subset of the K variables, would suffer from path dependence, as pointed out by Gelbach (2009). Despite these attractive properties, there are some important limitations to the standard OB decomposition that we now address in more detail.

[23] As is common in the gender pay gap literature, we begin with the counterfactual that use group B (males) as the reference group. In column (3) of Table 3, we present the decomposition that corresponds to Eq. (15), that is uses group A (females) as the reference group.

[24] In particular, see the discussion of the case of *scalable* or *categorical* variables below.

Table 3 Gender wage gap: Oaxaca-Blinder decomposition results (NLSY, 2000).

Reference group	(1) Using male coef. from col. 2, Table 2		(2) Using male coef. from col. 4, Table 2		(3) Using female coef. from col. 3, Table 2		(4) Using weighted sum of col. 2 and 3, Table 2		(5) Using pooled coef. from col. 5, Table 2	
Unadjusted mean log wage gap: $E[\ln(w_m)] - E[\ln(w_f)]$	0.233	(0.015)	0.233	(0.015)	0.233	(0.015)	0.233	(0.015)	0.233	(0.015)
Composition effects attributable to										
Age, race, region, etc.	0.012	(0.003)	0.012	(0.003)	0.009	(0.003)	0.011	(0.003)	0.010	(0.003)
Education	−0.012	(0.006)	−0.012	(0.006)	−0.008	(0.004)	−0.010	(0.005)	−0.010	(0.005)
AFQT	0.011	(0.003)	0.011	(0.003)	0.011	(0.003)	0.011	(0.003)	0.011	(0.003)
L.T. withdrawal due to family	0.033	(0.011)	0.033	(0.011)	0.035	(0.008)	0.034	(0.007)	0.028	(0.007)
Life-time work experience	0.137	(0.011)	0.137	(0.011)	0.087	(0.010)	0.112	(0.008)	0.092	(0.007)
Industrial sectors	0.017	(0.006)	0.017	(0.006)	0.003	(0.005)	0.010	(0.004)	0.009	(0.004)
Total explained by model	0.197	(0.018)	0.197	(0.018)	0.136	(0.014)	0.167	(0.013)	0.142	(0.012)

(*continued on next page*)

Table 3 (continued)

Reference group	(1) Using male coef. from col. 2, Table 2		(2) Using male coef. from col. 4, Table 2		(3) Using female coef. from col. 3, Table 2		(4) Using weighted sum of col. 2 and 3, Table 2		(5) Using pooled coef. from col. 5, Table 2	
Wage structure effects attributable to										
Age, race, region, etc.	−0.098	(0.234)	−0.098	(0.234)	−0.096	(0.232)	−0.097	(0.233)	−0.097	(0.240)
Education	0.045	(0.034)	0.045	(0.034)	0.041	(0.033)	0.043	(0.034)	0.043	(0.031)
AFQT	0.003	(0.023)	0.003	(0.023)	0.003	(0.025)	0.003	(0.024)	0.002	(0.025)
L.T. withdrawal due to family	0.003	(0.017)	0.003	(0.017)	0.001	(0.004)	0.002	(0.011)	0.007	(0.010)
Life-time work experience	0.048	(0.062)	0.048	(0.062)	0.098	(0.067)	0.073	(0.064)	0.092	(0.065)
Industrial sectors	−0.092	(0.033)	0.014	(0.028)	−0.077	(0.029)	−0.085	(0.031)	−0.084	(0.032)
Constant	0.128	(0.213)	0.022	(0.212)	0.193	(0.211)	0.128	(0.213)	0.128	(0.216)
Total wage structure/ Unexplained log wage gap	0.036	(0.019)	0.036	(0.019)	0.097	(0.016)	0.066	(0.015)	0.092	(0.014)

The data is an extract from the NLSY79 used in O'Neill and O'Neill (2006). The other explanatory variables are age, dummies for black, hispanic, region, msa, central city. In column (1), the omitted industrial sector is "Primary, Construction, and Utilities". In column (2), the omitted industrial sector is "Education, Health and Public Admin". Standard errors are in parentheses. The means of the variables are reported in Table 2.

3.2. Issues with detailed decompositions: choice of the omitted group

There are many relevant economic questions that can be answered with the detailed decomposition of the composition effect $\widehat{\Delta}_X^{\mu}$ in Eq. (18). For example, what has been the contribution of the gender convergence in college enrollment to the gender convergence in average pay? There are also some important questions that are based on the detailed decomposition of the wage structure effect $\widehat{\Delta}_S^{\mu}$. For example, consider the related "swimming upstream" query of Blau and Kahn (1997). To what extent have the increases in the returns to college slowed down the gender convergence in average pay? Or, to what extent has the decline in manufacturing and differences in industry wage premia contributed to that convergence?

Some difficulties of interpretation arise when the explanatory variables of interest are categorical (with more than two categories, or more generally, in the case of scalable variables, such as test scores) and do not have an absolute interpretation. In OB decompositions, categorical variables generate two problems. The first problem is that categorical or scalable variables do not have a natural zero, thus the reference point has to be chosen arbitrarily. The conventional practice is to omit one category which becomes the reference point for the other groups. This generates some interpretation issues even in the detailed decomposition of the composition effect.

Returning to our NLSY example, assume that the industry effects can captured by four dummy variables, $ind1$ to $ind4$, for the broad sectors: (i) primary, construction, transportation & utilities, (ii) manufacturing, (iii) education and health services & public administration, and (iv) other services. Consider the case where $ind1$ is the omitted category, $\beta_{g,ind1} = 0$, and denote by $\beta_{g,indk}$ the coefficients from the wage regression, as in column (2) of Table 2. Denote by $\beta'_{g,indk}$ the coefficients of a wage regression where $ind3$ is the omitted category, $\beta'_{g,ind3} = 0$, as in column (4) of Table 2, so that, for example, $\widehat{\beta}'_{g,ind4} = \widehat{\beta}_{g,ind4} - \widehat{\beta}_{g,ind3}[0.066 = 0.007 - (-0.059)]$. In our example, given the large difference in the coefficients of manufacturing between columns (2) and (4) of Table 2, this could mistakenly lead one to conclude that the effect of the underrepresentation of women in the manufacturing sector has an effect three times as large $(0.237 - 0.120) \times 0.093$ in one case (education and health omitted) as $(0.237 - 0.120) \times 0.034$ in the other case (primary omitted). In the first case, the underrepresentation of women in the manufacturing sector is 'priced' at the relative returns in the manufacturing versus the education and health sector, while in the other it is 'priced' at the relative returns in the manufacturing versus the primary sector.[25]

Note, however, that the overall effect of 0.017 (0.006) of gender differences in industrial sectors on the gender wage gap, is the same in columns (1) and (2) of Table 3.

[25] This interpretation issue also arises in other applications that use categorical variables, notably the inter-industry wage differentials literature. In this literature, following the seminal Krueger and Summers (1988) paper on inter-industry wage differentials, the standard practice is to express industry differentials as deviations from an employment-share weighted mean, a well-defined average.

To simplify the exposition, consider the special case where industrial sectors are the only explanatory factors in the wage regression. It follows that the composition effect,

$$\widehat{\Delta}_X^{\mu} = \sum_{k=1}^{4} \left[\overline{X}_{B,indk} - \overline{X}_{A,indk}\right] \widehat{\beta}_{A,indk}, \tag{19}$$

is unaffected by the choice of omitted category.[26]

The second problem with the conventional practice of omitting one category to identify the coefficients of the remaining categories is that in the unexplained part of the decomposition one cannot distinguish the part attributed to the group membership (true "unexplained" captured by the difference in intercepts) from the part attributed to differences in the coefficient of the omitted or base category.[27] These difficulties with the detailed decomposition of the unexplained part component were initially pointed by Jones (1983) who argued that "this latter decomposition is in most applications arbitrary and uninterpretable" (p. 126). Pursuing the example above, the effect of industry wage differentials on the gender wage gap is given by the right-hand side sums in the following expressions

$$\widehat{\Delta}_S^{\mu} = \left[(\widehat{\beta}'_{B0} + \widehat{\beta}'_{B,ind1}) - (\widehat{\beta}'_{A0} + \widehat{\beta}'_{A,ind1})\right] + \sum_{k\neq 1} \overline{X}_{B,indk}(\widehat{\beta}_{B,indk} - \widehat{\beta}_{A,indk}), \tag{20}$$

$$\widehat{\Delta}_S^{\mu'} = \left[(\widehat{\beta}_{B0} + \widehat{\beta}_{B,ind3}) - (\widehat{\beta}_{A0} + \widehat{\beta}_{A,ind3})\right] + \sum_{k\neq 3} \overline{X}_{B,indk}(\widehat{\beta}'_{B,indk} - \widehat{\beta}'_{A,indk}), \tag{21}$$

where $\widehat{\beta}_{g0} = \widehat{\beta}'_{g0} + \widehat{\beta}'_{g,ind1}$ and $\widehat{\beta}'_{g0} = \widehat{\beta}_{g0} + \widehat{\beta}_{g,ind3}$, $g = A, B$. The overall wage structure effect is the same irrespective of the omitted category $\widehat{\Delta}_S^{\mu} = \widehat{\Delta}_S^{\mu'}$, as shown in the last row of column (1) and (2) of Table 3. However, the overall effect of differences in the returns to industrial sectors, given by the right hand side sums with either choice of omitted group, −0.092 (0.033) in column (1) and 0.014 (0.028) in column (2), are different because different parts of the effect are hidden in the intercepts [0.128 (0.213) in column (1) and 0.022 (0.212) in column (2)].[28]

This invariance issue has been discussed by Oaxaca and Ransom (1999), Gardeazabal and Ugidos (2004), and Yun (2005, 2008), who have proposed tentative solutions to it.

[26] In the first regression, the composition effect is given by $\sum_{k\neq 1}(\overline{X}_{B,indk} - \overline{X}_{A,indk})\widehat{\beta}_{A,indk}$, and in the second regression, $\sum_{k\neq 3}(\overline{X}_{B,indk} - \overline{X}_{A,indk})\widehat{\beta}'_{A,indk} = \sum_{k\neq 3}(\overline{X}_{B,indk} - \overline{X}_{B,indk})[\widehat{\beta}_{A,indk} - \widehat{\beta}_{A,ind1}] = \sum_{k\neq 1}(\overline{X}_{A,indk} - \overline{X}_{A,indk})\widehat{\beta}_{A,indk}$ because $\sum_{k\neq 3}\overline{X}_{g,indk} = 1 - \overline{X}_{B,ind1}$, $g = A, B$.

[27] Actually, problems arise when they are more than two categories. Blinder (1973, footnote 13) and Oaxaca (2007) correctly point out that in the case of a binary dummy variable, these problems do not occur.

[28] This problem is different from a "true" identification problem which arises when multiple values of a parameter of interest are consistent with a given model and population.

These solutions impose some normalizations on the coefficients to purge the intercept from the effect of the omitted category, either by transforming the dummy variables before the estimation, or by implementing the restriction, $\sum_k \beta_{g,indk} = 0$, $g = A, B$, via restricted least squares.[29] Yun (2005) imposes the constraint that the coefficient on the first category equals the unweighted average of the coefficients on the other categories, $\beta_{g,ind1} = -\sum_{k\neq 1} \beta_{g,indk}/K$ along with $\sum_{k=1}^{K} \beta_{g,indk} = 0$. While these restrictions may appear to solve the problem of the omitted group, as pointed out by Yun (2008) "some degree of arbitrariness in deriving a normalized equation is unavoidable" (p. 31). For example, an alternative restriction on the coefficients, that goes back to Kennedy (1986), could be a weighted sum, $\sum_k w_k \beta_{gk} = 0$, where the weights w_k reflect the relative frequencies of the categories in the pooled sample. The coefficients would then reflect deviations from the overall sample mean.

The pitfall here is that the normalizations proposed by Gardeazabal and Ugidos (2004) and Yun (2005) may actually leave the estimation and decomposition without a simple meaningful interpretation. Moreover, these normalizations will likely be sample specific and preclude comparisons across studies. By contrast, in the case of educational categories, the common practice of using high school graduates as the omitted category allows the comparison of detailed decomposition results when this omitted category is comparable across studies.

Invariance of the detailed decomposition with respect to the choice of omitted category may appear to be a desirable property, but it is actually elusive and should not come at the expense of interpretability. There is no quick fix to the difficult choice of the appropriate omitted category or base group, which is actually exacerbated in procedures that go beyond the mean. To mimic the case of continuous variables, one may argue that an education category such as less than high school that yields the smallest wage effect should be the omitted one, but this category may vary more across studies than the high school category. Issues of internal logic have to be balanced with comparability across studies.

Another way of reporting the results of counterfactual experiments, proposed in the context of the gender wage gap by industry, is to report the wage structure effects for each k category by setting $\overline{X}_{g,indk} = 1$ and $\overline{X}_{g,indl} = 0$ for $l \neq k$ in the expression (20) for the total wage structure effect

$$\widehat{\delta}_g^{\mu}(indk) = (\widehat{\beta}_{B0} - \widehat{\beta}_{A0}) + \left(\widehat{\beta}_{B,indk} - \widehat{\beta}_{A,indk}\right) + \sum_{j=1}^{J} \overline{X}_{gj} \left(\widehat{\beta}_{Bj} - \widehat{\beta}_{Aj}\right)$$
$$k = 1, \ldots, 4, \tag{22}$$

[29] As pointed by Gardeazabal and Ugidos (2004), such restrictions can have some disturbing implications. In the case of educational categories, it rules out an outcome where group B members would earn higher returns than group A members for all levels of education.

in a case where there are other explanatory variables, X_j, $j = 1, J$.[30] Initially, such expressions included only the first two terms, the intercept and the effect of the category k (Fields and Wolff, 1995). Later, Horrace and Oaxaca (2001) added the wage structure effect associated with the other variables. This allows one to compare the effect of wage structure on gender wage differentials by category while controlling for other explanatory variables X_j, $j = 1, \ldots, J$ in a way that is invariant to the choice of omitted category.[31] In columns (1) and (2) of Table 3, the wage structure effect associated with variables other than industrial sectors is essentially zero, and the $\widehat{\delta}_A^{\mu}(indk)$ can be computed as the difference between the male and female coefficients in columns (2) and (3) of Table 2 plus the 0.128 difference in the constant, yielding values of 0.128, 0.022, 0.004, and 0.048 for industries 1 through 4, respectively. Horrace and Oaxaca (2001) also proposed to ex-post normalize the effects of each category with respect to the maximum categorical effect.

One disadvantage of decomposition terms like $\widehat{\delta}_g^{\mu}(indk)$ relative to the usual components of the detailed decomposition is that they do not sum up to the overall wage structure effect. As a result, just looking at the magnitude of the $\widehat{\delta}_g^{\mu}(indk)$ terms gives little indication of their quantitative importance in the decomposition. We propose a normalization to help assess the proportion of the total wage structure effect which can be attributed to a category k given that a proportion $\overline{X}_{g,indk}$ of group g workers belongs to that category, and that is also invariant to the choice of omitted category. The normalization uses the fact that the weighted sum of the $\widehat{\delta}_g^{\mu}(indk)$, $k = 1, \ldots, 4$ (that is, including the omitted category), is equal to the total wage structure effect, so that the proportional effect $\%\widehat{\delta}_{S,g}^{\mu}(indk)$ of category k in the total wage structure can be computed as[32]

$$\%\widehat{\delta}_S^{\mu}(indk) = \frac{\widehat{\delta}_g^{\mu}(indk)\overline{X}_{g,indk}}{\widehat{\Delta}_S^{\mu}} \quad \text{because} \quad \widehat{\Delta}_S^{\mu} = \sum_{k=1}^{4} \widehat{\delta}_g^{\mu}(indk)\overline{X}_{g,indk}. \tag{23}$$

In our empirical example, with group B as the reference group, this expression is computed using female averages, thus $\%\widehat{\delta}_S^{\mu}(indk)$ will tell us the proportion of the total wage structure effect that can be attributed to industrial category k given the proportion of women in each category. The numbers are 0.308 for primary, 0.074 for manufacturing, 0.040 for education and health, and 0.578 for other services. Despite being underrepresented in the manufacturing sector, because women's returns to manufacturing jobs are

[30] In the gender wage gap literature, when the reference wage structure is the male wage structure (group B) the means among women $\overline{X}_{Aj}$ will be used in Eq. (22).

[31] It is indeed easy to see that $\widehat{\delta}_g^{\mu'}(indk) = [(\widehat{\beta}_{B0} + \widehat{\beta}_{B,ind1}) - (\widehat{\beta}_{A0} + \widehat{\beta}_{A,ind1})] + [(\widehat{\beta}_{B,indk} - \widehat{\beta}_{B,ind1}) - (\widehat{\beta}_{A,indk} - \widehat{\beta}_{A,ind1})] + \sum_{j=1}^{J} \overline{X}_{gj}\left(\widehat{\beta}_{Bj} - \widehat{\beta}_{Aj}\right) = \widehat{\delta}^{\mu}(indk)$.

[32] The $\widehat{\delta}_g^{\mu}(indk)$ for the omitted category is simply the first and last components of Eq. (22), since $(\widehat{\beta}_{B,indk} - \widehat{\beta}_{A,indk}) = 0$ for that category.

relatively high, the share of the unexplained gap attributable to that factor turns out not to be that large.

3.3. Alternative choices of counterfactual

On the one hand, the choice of a simple counterfactual treatment is attractive because it allows us to use the identification results from the treatment effects literature. On the other hand, these simple counterfactuals may not always be appropriate for answering the economic question of interest. For instance, the male wage structure may not represent the appropriate counterfactual for the way women would be paid in the absence of labor market discrimination. If the simple counterfactual does not represent the appropriate treatment, it may be more appropriate to posit a new wage structure. For example, in the case of the gender pay gap, typically propositions (Reimers, 1983; Cotton, 1998; Neumark, 1988; Oaxaca and Ransom, 1994) have used a weighted average expression $\beta^* = \Omega\beta_A + (I - \Omega)\beta_B$, where $\Omega = I$ corresponds to $\beta^* = \beta_A$, $\Omega = 0$ corresponds to $\beta^* = \beta_B$, and where $\Omega = \omega \cdot I$ could reflect a weighting corresponding to the share of the two groups in the population. Another popular choice is the matrix $\Omega^* = (\mathbf{X}_B^\intercal \mathbf{X}_B + \mathbf{X}_A^\intercal \mathbf{X}_A)^{-1} \mathbf{X}_B^\intercal \mathbf{X}_B$, which captures the sample variation in the characteristics of group A and B workers.[33] The decomposition is then based on the triple differences:

$$\begin{aligned}\widehat{\Delta}_O^\mu &= \left(\overline{X}_B - \overline{X}_A\right)\widehat{\beta}^* + \left[\overline{X}_B\left(\widehat{\beta}_B - \widehat{\beta}^*\right) + \overline{X}_A\left(\widehat{\beta}^* - \widehat{\beta}_A\right)\right] \\ &= \widehat{\Delta}_X^\mu + \widehat{\Delta}_S^\mu.\end{aligned}$$

Table 3 shows that in the NLSY example, the gender gap decomposition is substantially different when either the female wage structure (column 3) or the weighted sum of the male and female wage structure (column 4) is used as the reference wage structure. Typically (as in Bertrand and Hallock (2001) for example), with the female wage structure as reference, the explained part of the decomposition (composition effect) is smaller than with the male wage structure as reference. Indeed, evaluated at either female 'prices' or average of male and female 'prices', the total unexplained (wage structure) effect becomes statistically significant.

An alternative measure of "unexplained" differences (see Cain, 1986) in mean wages between group A and group B workers is given by the coefficient δ of the group membership indicator variable D_B in the wage regression on the pooled sample, where the coefficients of the observed wage determination characteristics are constrained to be the same for both groups:

$$\mathbb{E}[Y_i|X, D_B] = \alpha_0 + X_i\beta^{**} + \delta D_{Bi}, \tag{24}$$

[33] $\mathbf{X}_A$ and $\mathbf{X}_B$ are the matrices of covariates (of dimension $N_A \times k$ and $N_B \times k$) for groups A and B, respectively.

where the vector of observed characteristics X_i excludes the constant. It follows that,

$$\begin{aligned}\Delta_O^{\mu} &= \mathbb{E}[Y_i|X, D_B = 1] - \mathbb{E}[Y_i|X, D_B = 0] \\ &= \left(\alpha_0 + \mathbb{E}[X_i|D_B = 1]\beta^{**} + \delta\right) - \left(\alpha_0 + \mathbb{E}[X_i|D_B = 0]\beta^{**}\right) \\ &= \left(\mathbb{E}[X_i|D_B = 1] - \mathbb{E}[X_i|D_B = 0]\right)\beta^{**} + \delta = \Delta_X^{\mu} + \Delta_S^{\mu},\end{aligned}$$

where $\delta = \Delta_S^{\mu}$. As noted by Fortin (2008), this "regression-compatible" approach is preferable to the one based on a pooled regression that omits the group membership variable (as in Neumark (1988) and Oaxaca and Ransom (1994)), because in the latter case the estimated coefficients are biased (omitted variable bias). Note, however, that this counterfactual corresponds to the case where the group membership dummy is thought to be sufficient to purge the reference wage structure from any group membership effect, an assumption that is maintained in the common practice of using the group membership dummy in a simple regression to assess its effect. The detailed decomposition is obtained using the above triple differences decomposition.[34]

The results of this decomposition, reported in column (5) of Table 3, are found to be closest to the one using the female coefficients in column (3), but this is not necessarily always the case. Notice that the magnitude of the total unexplained wage log wage gap 0.092 (0.014) log points corresponds to the coefficient of the female dummy in column (5) of Table 2.

3.4. Reweighted-regression decompositions

A limitation of OB decompositions, discussed by Barsky et al. (2002), is that they may not provide consistent estimates of the wage structure and composition effect when the conditional mean function is non linear. Barsky et al. (2002) look at the role of earnings and other factors in the racial wealth gap. They argue that a standard OB decomposition is inadequate because the wealth-earnings relationship is non linear, and propose a more flexible approach instead.

Under the linearity assumption, the average counterfactual wage that group B workers would have earned under the wage structure of group A is equal to $\mathbb{E}[X_B|D_B = 1] \cdot \beta_A$, and is estimated as the product $\overline{X}_B\widehat{\beta}_A$, a term that appears in both the wage structure and composition effect in Eq. (15). However, when linearity does not hold, the counterfactual mean wage will not be equal to this term.

One possible solution to the problem is to estimate the conditional expectation using non-parametric methods. Another solution proposed by Barsky et al. (2002) is to use a (non-parametric) reweighting approach as in DiNardo et al. (1996) to perform the decomposition. One drawback of this decomposition method, discussed later in the

[34] This "pooled" decomposition is easily implemented using the option "pooled" in Jann (2008) "oaxaca" procedure in Stata 9.2.

chapter, is that it does not provide, in general, a simple way of performing a detailed decomposition. In the case of the mean, however, this drawback can be readily addressed by estimating a regression in the reweighted sample.

To see this, let $\Psi(X_i)$ be the reweighting function, discussed in Section 4.5, that makes the characteristics of group A workers similar to those of group B workers. The counterfactual coefficients β_A^C and the counterfactual mean $\overline{X}_A^C$, are then estimated as:[35]

$$\widehat{\beta}_A^C = \left(\sum_{i\in A} \widehat{\Psi}(X_i)\cdot X_i \cdot X_i^{\intercal}\right)^{-1} \cdot \sum_{i\in A} \widehat{\Psi}(X_i)\cdot Y_{Ai}\cdot X_i$$
$$\overline{X}_A^C = \sum_{i\in A} \widehat{\Psi}(X_i)\cdot X_i,$$

where $plim(\overline{X}_A^C) = plim(\overline{X}_B) = \mathbb{E}(X|D_B = 1)$.[36] If the conditional expectation of Y given X was linear, both the weighted and unweighted regressions would yield the same consistent estimate of β_A, i.e. we would have $plim(\widehat{\beta}_A^C) = plim(\widehat{\beta}_A) = \beta_A$. When the conditional expectation is not linear, however, the weighted and unweighted estimates of β_A generally differ since OLS minimizes specification errors over different samples.[37]

Consider the "reweighted-regression" decomposition of the overall wage gap $\widehat{\Delta}_{O,R}^{\mu}$, where

$$\begin{aligned}\widehat{\Delta}_{O,R}^{\mu} &= (\overline{X}_B\widehat{\beta}_B - \overline{X}_A^C\widehat{\beta}_A^C) + (\overline{X}_A^C\widehat{\beta}_A^C - \overline{X}_A\widehat{\beta}_A)\\ &= \widehat{\Delta}_{S,R}^{\mu} + \widehat{\Delta}_{X,R}^{\mu}.\end{aligned}$$

The composition effect $\widehat{\Delta}_{X,R}^{\mu}$ can be divided into a pure composition effect $\widehat{\Delta}_{X,p}^{\mu}$ using the wage structure of group A, and a component linked to the specification error

35 When considering covariates X, we use the subscript g to denote the group whose characteristics are "adjusted" with reweighting.

36 We show in Section 4 that the reweighting factor $\Psi(X)$ is defined as the ratio of the marginal distributions of X for groups B and A, $\Psi(X) = \mathrm{d}F_{X_B}(X)/\mathrm{d}F_{X_A}(X)$. As a result, the reweighted distribution of X for group A should be the same as the original distribution of X in group B. This implies that the mean value of X in the reweighted sample, $\overline{X}_A^C$, should be the same as the mean value of X for group B, $\overline{X}_B$.

37 When the conditional expectation is non-linear, the OLS estimate of β can be interpreted as the one which minimizes the square of the specification error $\mathbb{E}(Y|X) - X\beta$ over the distribution of X. Since the expected value of the OLS estimate of β depends on the distribution of X, differences in β over two samples may either reflect true underlying differences in the conditional expectation (i.e. in the wage structure), or "spurious" differences linked to the fact that the distribution of X is different in the two samples. For example, if $\mathbb{E}(Y|X)$ is convex in X, the expected value of β will tend to grow as the distribution of X shifts up, since the relationship between Y and X gets steeper as X becomes larger.

in the linear model, $\widehat{\Delta}^{\mu}_{X,e}$:

$$\begin{aligned}\widehat{\Delta}^{\mu}_{X,R} &= \left(\overline{X}^{C}_{A} - \overline{X}_{A}\right)\widehat{\beta}_{A} + \overline{X}^{C}_{A}\left[\widehat{\beta}^{C}_{A} - \widehat{\beta}_{A}\right] \\ &= \widehat{\Delta}^{\mu}_{X,p} + \widehat{\Delta}^{\mu}_{X,e}.\end{aligned}$$

The wage structure effect can be written as

$$\begin{aligned}\widehat{\Delta}^{\mu}_{S,R} &= \overline{X}_{B}\left(\widehat{\beta}_{B} - \widehat{\beta}^{C}_{A}\right) + \left(\overline{X}_{B} - \overline{X}^{C}_{A}\right)\widehat{\beta^{C}_{A}} \\ &= \widehat{\Delta}^{\mu}_{S,p} + \widehat{\Delta}^{\mu}_{S,e}\end{aligned}$$

and reduces to the first term $\widehat{\Delta}^{\mu}_{S,p}$ as the reweighting error $\widehat{\Delta}^{\mu}_{S,e}$ goes to zero in large samples ($plim(\overline{X}_{B} - \overline{X}^{C}_{A}) = 0 \Rightarrow plim(\widehat{\Delta}^{\mu}_{S,e}) = 0$).

The reweighted-regression decomposition is similar to the usual OB decomposition except for two small differences. The first difference is that the wage structure effect is based on a comparison between $\widehat{\beta}_{B}$ and the weighted estimate $\widehat{\beta}^{C}_{A}$ instead of the usual unweighted estimate $\widehat{\beta}_{A}$. As discussed in Firpo et al. (2007), this ensures that the difference $\widehat{\beta}_{B} - \widehat{\beta}^{C}_{A}$ reflects true underlying differences in the wage structure for group A and B, as opposed to a misspecification error linked to the fact that the underlying conditional expectation is non-linear. Note that is also useful to check whether the reweighting error $(\overline{X}_{B} - \overline{X}^{C}_{A})\widehat{\beta}^{C}_{A}$ is equal to zero (or close to zero), as it should be when the reweighting factor $\widehat{\Psi}(X)$ is consistently estimated.

The other difference relative to the OB decomposition is that the composition effects consists of a standard term $(\overline{X}^{C}_{A} - \overline{X}_{A})\widehat{\beta}_{A}$ plus the specification error $\overline{X}^{C}_{A}\left[\widehat{\beta}^{C}_{A} - \widehat{\beta}_{A}\right]$. If the model was truly linear, the specification error term would be equal to zero. Computing the specification error is important, therefore, for checking whether the linear model is well specified, and adjusting the composition effect in the case where the linear specification is found to be inaccurate.

In the case where the conditional expectation $\mathbb{E}(Y_i|X_i, D = d)$ is estimated non-parametrically, a whole different procedure would have to be used to separate the wage structure into the contribution of each covariate. For instance, average derivative methods could be used to estimate an effect akin to the β coefficients used in standard decompositions. Unfortunately, these methods are difficult to use in practice, and would not be helpful in dividing up the composition effect into the contribution of each individual covariate.

On a related note, Kline (2009) points out that the standard OB decomposition can be interpreted as a reweighting estimator where the weights have been linearized as a function of the covariates. This suggests that the procedure may actually be more robust to departures from linearity than what has been suggested in the existing literature. Since

the procedure is robust to these departures and remains the method of choice when linearity holds, Kline (2009) points out that it is "doubly robust" in the sense of Robins et al. (1994) and Egel et al. (2009).

3.5. Extensions to limited dependent variable models

OB decompositions have been extended to cases where the outcome variable is not a continuous variable. To mention a few examples, Gomulka and Stern (1990) study the changes over time in labor force participation of women in the United Kingdom using a probit model. Even and Macpherson (1990) decomposes the male-female difference in the average probability of unionization, while Doiron and Riddell (1994) propose a decomposition of the gender gap in unionization rate based on a first order Taylor series approximation of the probability of unionization. Fitzenberger et al. (forthcoming) use a probit model to decompose changes over time in the rate of unionization in West and East Germany. Fairlie (1999, 2005) discuss the cases of the racial gaps in self-employment and computer ownership. Bauer and Sinning (2008) discuss the more complicated cases of a count data model, for example where the dependent variable is the number of cigarettes smoked by men and women (Bauer et al., 2007), and of the truncated dependent variable, where for example the outcome of interest is hours of work.

In the case of a limited dependent variable Y, the conditional expectation of Y is typically modeled as a non-linear function in X, $\mathbb{E}(Y_g|X;\beta_g) = G(X;\beta_g)$. For example, if Y is a dichotomous outcome variable ($Y = 0, 1$) and $Y_g^* = X\beta_g + \upsilon_g$ is a latent variable which is linear in X, it follows that $\mathbb{E}(Y_g|X;\beta_g) = G(X\beta_g)$ where $G(\cdot)$ is the CDF of υ_g. When υ_g follows a standard normal distribution, we have a standard probit model and $G(\cdot) = \Phi(\cdot)$. More generally, under various assumptions regarding the functional form G and/or the distribution of the error terms υ_g, the models are estimated by maximum likelihood.

Because $\mathbb{E}(Y_g|D_g = 1) = \mathbb{E}[\mathbb{E}(Y_g|X;\beta_g)|D_g = 1] = \mathbb{E}[G(X;\beta_g)|D_g = 1] \neq G(\mathbb{E}[X|D_g = 1];\beta_g)$, the decomposition cannot simply be computed by plugging in the estimated β's and the mean values of X's, as in the standard OB decomposition. Counterfactual conditional expectations have to be computed instead, and averaged across observations. For example, if group A is thought to be the reference group, $\mathbb{E}(Y_B|D_A = 1) = \mathbb{E}[G(X;\beta_B)|D_A = 1]$ will be the counterfactual conditional expectation of Y_B that would prevail if the coefficients of the determinants of self-employment (for example) for group B were the same as for group A. This involves computing predicted (i.e. expected) values based on the estimated model for group B, $G(X;\beta_B)$, over all observations in group A, and averaging over these predicted values.

The mean gap between group B and group A is then decomposed as follows

$$\begin{aligned}\Delta_O^{\mu} &= \mathbb{E}[(Y_B|D_B = 1)] - \mathbb{E}(Y_A|D_A = 1)\\ &= \mathbb{E}[G(X;\beta_B)|D_B = 1] - \mathbb{E}[G(X;\beta_A)|D_A = 1]\end{aligned}$$

$$
\begin{aligned}
&= (\mathbb{E}[G(X;\beta_B)|D_B = 1] - \mathbb{E}[G(X;\beta_A)|D_B = 1]) \\
&\quad + (\mathbb{E}[G(X;\beta_A)|D_B = 1] - \mathbb{E}[G(X;\beta_A)|D_A = 1]) \\
&= \Delta_S^{\mu} + \Delta_X^{\mu},
\end{aligned}
$$

into a component that attributes differences in the mean outcome variable to differences in the characteristics of the individuals, and a component that attributes these differences to differences in the coefficients.

The same difficult issues in the appropriate choice of counterfactuals persist for more general non-linear models. In addition, extra care has to be taken to verify that the sample counterfactual conditional expectation lies within the bounds of the limited dependent variable. For example, Fairlie (1999) checks that average self-employment for Blacks predicted from the White coefficients is not negative.

The non-linear decomposition may perform better than the linear alternative (linear probability model, LPM) when the gap is located in the tails of the distribution or when there are very large differences in the explanatory variables, whose effects would remain unbounded in a LPM. On the other hand, there are many challenges in the computation of detailed decompositions for non-linear models. Because of non-linearity, the detailed decomposition of the two components into the contribution of each variable, even if the decomposition was linearized using marginal effects, would not add up to the total. Gomulka and Stern (1990) and Fairlie (2005) have proposed alternative methodologies based on a series of counterfactuals, where the coefficient of each variable is switched to reference group values in sequence. In the latter cases, the decomposition will be sensitive to the order of the decomposition, that is will be path dependent. We discuss these issues further in the context of the decompositions of entire distributions in Section 5.

3.6. Statistical inference

OB decompositions have long been presented without standard errors. More recently, Oaxaca and Ransom (1998), followed by Greene (2003, p. 53–54), have proposed approximate standard errors based the delta method, under the assumption that the explanatory variables were fixed.[38] A more modern approach where, as above, (Y, X) are stochastic was suggested and implemented by Jann (2005). In cases where the counterfactuals are not a simple treatment, or where a non-linear estimator is used, bootstrapping the entire procedure may prove to be the practical alternative.

4. GOING BEYOND THE MEAN—DISTRIBUTIONAL METHODS

Developing new decomposition methods for distributional statistics other than the mean has been an active research area over the last 15 years. In this section, we discuss a number

[38] This corresponds to an experimental setting where, for example, regression analysis was used to assess the impact of various soils and fertilizers (X) on agricultural yields Y.

of procedures that have been suggested for decomposing general distributional statistics. We focus on the case of the aggregate decomposition, though some of the suggested methods can be extended to the case of the detailed decomposition, which we discuss in Section 5. We begin by looking at the simpler case of a variance decomposition. The decomposition is obtained by extending the classic analysis of variance approach (based on a between/within group approach) to a general case with covariates X. We then turn to new approaches based on various "plugging in" methods such as JMP's residual imputation method and Machado and Mata (2005)'s conditional quantile regression method. Finally, we discuss methods that focus on the estimation of counterfactuals for the entire distribution. These methods are either based on reweighting or on the estimation of the conditional distribution.

Most of this recent research was initially motivated by the dramatic growth in earnings inequality in the United States. Prior to that episode, the literature was considering particular summary measures of inequality such as the variance of logs and the Gini coefficient. For instance, Freeman (1980, 1984) looks at the variance of log wages in his influential work on the effect of unions on wage dispersion. This research establishes that unions tend to reduce wage dispersion as measured by the variance of log wages. Freeman shows that despite the inequality-enhancing effect of unions on the between-group component of inequality, the overall effect of unions is to reduce inequality because of the even larger effect of unions on within-group inequality.

One convenient feature of the variance is that it can be readily decomposed into a within- and between-group component. Interestingly, related work in the inequality literature shows that other measures such as the Gini or Theil coefficient are also decomposable into a within- and between-group component.[39]

Note that the between vs. within decomposition is quite different in spirit from the aggregate or detailed OB decomposition discussed in the previous section. There are advantages and disadvantages to this alternative approach. On the positive side, looking at between- and within-group effects can help understand economic mechanisms, as in the case of unions, or the sources of inequality growth (Juhn et al., 1993).

On the negative side, the most important drawback of the between vs. within decomposition is that it does not hold in the case of many other interesting inequality measures such as the interquartile ranges, the probability density function, etc. This is a major shortcoming since looking at what happens where in the distribution is important for identifying the factors behind changes or differences in distributions. Another drawback of the between vs. within approach is that it does not provide a straightforward way of looking at the specific contribution of each covariate, i.e. to perform a detailed decomposition. One final drawback is that with a rich enough set of covariates the number of possible groups becomes very large, and some parametric restrictions have to be introduced to keep the estimation problem manageable.

[39] See, for instance, Bourguignon (1979), Cowell (1980), and Shorrocks (1980, 1984).

In response to these drawbacks, a new set of approaches have been proposed for performing aggregate decompositions on any distributional statistic. Some approaches such as Juhn et al. (1993), Donald et al. (2000), and Machado and Mata (2005) can be viewed as extensions of the variance decomposition approach where the whole conditional distribution (instead of just the conditional variance) are estimated using parametric approaches. Others such as DiNardo et al. (1996) completely bypass the problem of estimating conditional distributions and are, as such, closer cousins to estimators proposed in the program evaluation literature.

4.1. Variance decompositions

Before considering more general distributional statistics, it is useful to recall the steps used to obtain the standard OB decomposition. The first step is to assume that the conditional expectation of Y given X is linear, i.e. $\mathbb{E}(Y|X) = X\beta$. This follows directly from the linearity and zero conditional mean assumptions (Assumptions 10 and 11) introduced in Section 2. Using the law of conditional expectations, it then follows that the unconditional mean is $\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(X)\beta$. This particular property of the mean is then used to compute the OB decomposition.

In light of this, it is natural to think of extending this type of procedure to the case of the variance. Using the analysis of variance formula, the unconditional variance of Y can be written as:[40]

$$\begin{aligned}\text{Var}(Y) &= \mathbb{E}[\text{Var}(Y|X)] + \mathbb{E}\left\{[\mathbb{E}(Y|X) - \mathbb{E}(Y)]^2\right\} \\ &= \mathbb{E}[\text{Var}(Y|X)] + \mathbb{E}\left\{[X\beta - \mathbb{E}(X)\beta]^2\right\} \\ &= \mathbb{E}[\text{Var}(Y|X)] + \beta'\text{Var}(X)\beta,\end{aligned}$$

where the expectations are taken over the distribution of X. The first component of the equation is the within-group component (also called residual variance), while the second component is the between-group component (also called regression variance). Writing $\text{Var}(Y|X, D_g = 1) \equiv \text{v}_g(X)$, $g = A, B$, we can write the difference in variances across groups B and A as

$$\begin{aligned}\Delta_O^V = {} & \mathbb{E}\left[\text{v}_B(X)|D_B = 1\right] - \mathbb{E}\left[\text{v}_A(X)|D_B = 0\right] + \beta_B'\text{Var}\left[X|D_B = 1\right]\beta_B \\ & - \beta_A'\text{Var}\left[X|D_B = 0\right]\beta_A.\end{aligned}$$

A few manipulations yield $\Delta_O^V = \Delta_X^V + \Delta_S^V$, where

$$\begin{aligned}\Delta_X^V = {} & \left\{\mathbb{E}\left[\text{v}_A(X)|D_B = 1\right] - \mathbb{E}\left[\text{v}_A(X)|D_B = 0\right]\right\} \\ & + \beta_A'\left\{\text{Var}\left[X|D_B = 1\right] - \text{Var}\left[X|D_B = 0\right]\right\}\beta_A\end{aligned}$$

[40] See for example, Theorem B.4 in Greene (2003).

and

$$\Delta_S^V = \{\mathbb{E}[\mathrm{v}_B(X)|D_B = 1] - \mathbb{E}[\mathrm{v}_A(X)|D_B = 1]\} + (\beta_B - \beta_A)'\mathrm{Var}[X|D_B = 1](\beta_B + \beta_A).$$

While it is straightforward to estimate the regression coefficients (β_A and β_B) and the covariance matrices of the covariates ($\mathrm{Var}[X|D_B = 0]$ and $\mathrm{Var}[X|D_B = 1]$), the within-group (or residual) variance terms $\mathrm{v}_A(X)$ and $\mathrm{v}_B(X)$ also have to be estimated to compute the decomposition.

Several approaches have been used in the literature to estimate $\mathrm{v}_A(X)$ and $\mathrm{v}_B(X)$. The simplest possible approach is to assume that the error term is homoscedastic, in which case $\mathrm{v}_A(X) = \sigma_A^2$ and $\mathrm{v}_B(X) = \sigma_B^2$, and the two relevant variance parameters can be estimated from the sampling variance of the error terms in the regressions. The homoscedasticity assumption is very strong, however. When errors are heteroscedastic, differences between σ_A^2 and σ_B^2 can reflect spurious composition effects, in which case the decomposition will attribute to the wage structure effect (Δ_S^V) what should really be a composition effect (Δ_X^V). Lemieux (2006b) has shown this was a major problem when looking at changes in residual wage inequality in the United States since the late 1980s.

A simple way of capturing at least some of the relationship between the covariates and the conditional variance is to compute the variance of residuals for a limited number of subgroups of "cells". For instance, Lemieux (2006b) shows estimates for 20 different subgroups of workers (based on education and experience), while Card (1996) divides the sample into five quintiles based on predicted wages $X\widehat{\beta}$.

Finally, one could attempt to estimate a more general specification for the conditional variance by running a "second step" model for squared regression residual $\widehat{\upsilon}^2 = \left(Y - X\widehat{\beta}\right)^2$ on some specification of the covariates. For example, assuming that $\mathrm{v}_A(X) = X\delta_A$, we can estimate $\widehat{\delta}$ by running a regression of $\widehat{\upsilon}^2$ on X.[41] We can then write the two aggregate components of the variance decomposition as:

$$\Delta_X^V = \{(\mathbb{E}[X|D_B = 1] - \mathbb{E}[X|D_B = 0])\,\delta_A\} + \beta_A'\{\mathrm{Var}[X|D_B = 1] - \mathrm{Var}[X|D_B = 0]\}\,\beta_A \quad (25)$$

and

$$\Delta_S^V = \{\mathbb{E}[X|D_B = 1](\delta_B - \delta_A)\} + (\beta_B - \beta_A)'\mathrm{Var}[X|D_B = 1](\beta_B + \beta_A). \quad (26)$$

Compared to the standard OB decomposition for the mean, which only requires estimating a (regression) model for the conditional mean, in the case of the variance, we

[41] Estimating these simple models of the conditional cross-sectional variance is a special case of the large time-series literature on the estimation of auto-regressive conditional heteroskedasticity models (ARCH, GARCH, etc.).

also need to estimate a model for the conditional variance. While this is quite feasible in practice, we can already see a number of challenges involved when decomposing distributional parameters beyond the mean:

- The estimation is more involved since we need to estimate models for two, instead of just one, conditional moment. Furthermore, little guidance is typically available on "reasonable" specifications for the conditional variance. For instance, in the case of wages, the Mincer equation provides a reasonably accurate and widely accepted specification for the conditional mean, while no such standard model is available for the conditional variance.
- Computing the detailed decomposition is more complicated since the between-group component is a quadratic form in the β's. This yields a number of interaction terms that are difficult to interpret.

Since the complexity of decomposition methods already increases for a distributional measure as simple and convenient as the variance, this suggests these problems will be compounded in the case of other distributional measures such as quantiles. Indeed, we show in the next subsection that for quantiles, attempts at generalizing the approach suggested here require estimating the entire conditional distribution of Y given X. This is a more daunting estimation challenge, and we now discuss solutions that have been suggested in the literature.

4.2. Going beyond the variance: general framework

An important limitation of summary measures of dispersion such as the variance, the Gini coefficient or the Theil coefficient is that they provide little information regarding what happens where in the distribution. This is an important shortcoming in the literature on changes in wage inequality where many important explanations of the observed changes have specific implications for specific points of the distribution. For instance, the minimum wage explanation suggested by DiNardo et al. (1996) should only affect the bottom end of the distribution. At the other extreme, explanations based on how top executives are compensated should only affect the top of the distribution. Other explanations based on de-unionization (Freeman, 1993; Card, 1992; DiNardo et al., 1996) and the computerization of "routine" jobs (Autor et al., 2003) tend to affect the middle (or "lower middle") of the distribution. As a result, it is imperative to go beyond summary measures such as the variance to better understand the sources of growing wage inequality.

Going beyond summary measures is also important in many other interesting economic problems such the sources of the gender wage gap and the impact of social programs on labor supply.[42] The most common approach for achieving this goal is to perform a decomposition for various quantiles (or differences between quantiles like the

[42] See Albrecht et al. (2003), who look at whether there is a glass ceiling in female earnings, and Bitler et al. (2006), who study the distributional effects of work incentive programs on labor supply.

90-10 gap) of the distribution. Unfortunately, as we point out in the introduction, it is much more difficult to decompose quantiles than the mean or even the variance. The basic problem is that the law of iterated expectations does not hold in the case of quantiles, i.e. $Q_{g,\tau} \neq \mathbb{E}_X[Q_{g,\tau}(X)]$, where $Q_{g,\tau}$, is the τth quantile of the (unconditional) distribution of Y_g, and $Q_{g,\tau}(X)$ is the corresponding conditional quantile.

As it turns out, one (implicitly) needs to know the entire conditional distribution of Y_g given X given to compute $Q_{g,\tau}$. To see this, note that

$$\tau = F_{Y_g}(Q_{g,\tau}) = \mathbb{E}[F_{Y_g|X_g}(Q_{g,\tau}|X)] = \int F_{Y_g|X_g}(Q_{g,\tau}|X)\mathrm{d}F_{X_g}(X), \quad g = A, B,$$

where $F_{Y_g|X_g}(\cdot)$ is the cumulative distribution of Y conditional on X in group g. Given τ, it is possible to implicitly use this equation to solve for $Q_{g,\tau}$. It is also clear that in order to do so we need to know the conditional distribution function $F_{Y_g|X_g}(\cdot)$, as opposed to just the conditional mean and variance, as was the case for the variance. Estimating an entire conditional distribution function for each value of $(Y_g|X)$ is a difficult problem. Various decomposition methods that we discuss in detail below suggest different ways of handling this challenge.

But before covering them in detail, we recall the basic principles underlying these methods. As in Section 2, we focus on cumulative distributions since any standard distribution statistic, such as a quantile, can be directly computed from the cumulative distribution. For instance, quantiles of the counterfactual distribution can be obtained by inverting $F_{Y_A^C}$: $Q^C_{A,\tau} = F^{-1}_{Y_A^C}(\tau)$.

For the sake of presentational simplicity, we introduce a simplified notation relative to Section 2. We use F_{X_g} instead of $F_{X|D_g}$ to represent the marginal distribution of X, and $F_{Y_g|X_g}$ to represent $F_{Y_g|X,D_g}$ the conditional distributions, for $g = A, B$, introduced in Eq. (4). We use the shorthand $F_{Y_A^C}$ instead of $F_{Y_A^C:X=X|D_B}$ to represent the key counterfactual distribution of interest introduced in Eq. (5), which mixes the distribution of characteristics of group B with the wage structure from group A:

$$F_{Y_A^C}(y) = \int F_{Y_A|X_A}(y|X)\mathrm{d}F_{X_B}(X). \tag{27}$$

Three general approaches have been suggested in the decomposition literature for estimating the counterfactual distribution $F_{Y_A^C}(y)$. A first general approach, initially suggested by Juhn et al. (1993), replaces each value of Y_B for group B with a counterfactual value of $Y_A^C = g(Y_B, X)$, where $g(\cdot,\cdot)$ is an imputation function. The idea is to replace Y_B from group B with a counterfactual value of Y_A^C that holds the same rank in the conditional distribution $F_{Y_A|X_A}(\cdot|\cdot)$ as it did in the original distribution of Y_B. As we discussed in Section 2.2.3, this is done in practice using a residual imputation procedure.

Machado and Mata (2005) and Autor et al. (2005) have later suggested other approaches, based on conditional quantile regressions, to transform a wage observation Y_B into a counterfactual observation Y_A^C.

A second approach proposed by DiNardo et al. (1996) [DFL] is based on the following manipulation of Eq. (27):

$$F_{Y_A^C}(y) = \int F_{Y_A|X_A}(y|X)\Psi(X)\mathrm{d}F_{X_A}(X), \tag{28}$$

where $\Psi(X) = \mathrm{d}F_{X_B}(X)/\mathrm{d}F_{X_A}(X)$ is a reweighting factor. This makes it clear that the counterfactual distribution $F_{Y_A^C}(\cdot)$ is simply a reweighted version of the distribution $F_{Y_A}(\cdot)$. The reweighting factor is a simple function of X that can be easily estimated using standard methods such as a logit or probit. The basic idea of the DFL approach is to start with group A, and then replace the distribution of X of group A ($F_{X_A}(\cdot)$) with the distribution of X of group B ($F_{X_B}(\cdot)$) using the reweighting factor $\Psi(\cdot)$.

The third set of approaches also works with Eq. (27) starting with group B, and then replacing the conditional distribution $F_{Y_B|X_B}(Y|X)$ with $F_{Y_A|X_A}(Y|X)$. Doing so is more involved, from an estimation point of view, than following the DFL approach. The problem is that the conditional distributions depend on both X and y, while the reweighting factor $\Psi(X)$ only depends on X.

Under this third set of approaches, one needs to directly estimate the conditional distribution $F_{Y|X}(y|X)$. Parametric approaches for doing so were suggested by Donald et al. (2000) who used a hazard model approach, and Fortin and Lemieux (1998) who suggested estimating an ordered probit. More recently, Chernozhukov et al. (2009) suggest estimating distributional regressions (e.g. a logit, for each value of y). In all cases, the idea is to replace the conditional distribution for group B, $F_{Y_B|X_B}(y|X)$, with an estimate of the conditional distribution $F_{Y_A|X_A}(y|X)$ obtained using one of these methods.

In the next subsections, we discuss how these various approaches can be implemented. We also present some results regarding their statistical properties, and address computational issues linked to their implementation.

4.3. Residual imputation approach: JMP

Procedure

As we explain above, Juhn et al. (1993) propose an imputation approach where the wage Y_B from group B is replaced by a counterfactual wage Y_A^C where both the returns to observables and unobservables are set to be as in group A. The implementation of this procedure is divided in two steps. First, unobservables are replaced by counterfactual unobservables, as in Eq. (9). Second, counterfactual returns to observables are also imputed, as in Eq. (12).[43]

[43] Juhn et al. (1993) actually consider multiple time periods and proposed an additional counterfactual where the returns to observables are set to their mean across time periods, a complex counterfactual treatment.

Under the assumption of additive linearity (Assumption 10), the original wage equation for individual i from group B,

$$Y_{Bi} = X_i\beta_B + \upsilon_{Bi} \quad \text{where } \upsilon_{Bi} = h_B(\varepsilon_i)$$

allows the returns to unobservables to be group-specific. Under the assumption of rank preservation (14), the first counterfactual is computed as

$$Y_{Ai}^{C,2} = X_i\beta_B + \upsilon_{Ai}^{C,2}, \tag{29}$$

where

$$\upsilon_{Ai}^{C,2} = F_{\upsilon_A|X}^{-1}(\tau_{Bi}(x_i), x_i),$$

and $\tau_{Bi}(x_i)$ is the conditional rank of υ_{Bi} in the distribution of residuals for group B ($\tau_{Bi}(x_i) = F_{\upsilon_B|X}(\upsilon_{Bi}|X = x_i)$). A second counterfactual is then obtained by also replacing the returns to observable characteristics β_B with β_A

$$Y_{Ai}^{C,3} = X_i\beta_A + \upsilon_{Ai}^{C,2}.$$

Under the assumptions of linearity and rank preservation, this counterfactual wage should be the same as Y_{Ai}^{C}, the counterfactual wage obtained by replacing the wage structure $m_B(\cdot)$ with $m_A(\cdot)$.

In practice, it is straightforward to estimate β_A and β_B using OLS under the assumptions of linearity and zero conditional mean. It is much less clear, however, how to perform the residual imputation procedure described above. Under the strong assumption that the regression residuals υ_g are independent of X, it follows that

$$\upsilon_{Ai}^{C,2} = F_{\upsilon_A}^{-1}(\tau_{Bi}).$$

Under this independence assumption, one simply needs to compute the rank of the residual υ_{Bi} in the marginal distribution (distribution over the whole sample) of residuals for group B, and then pick the corresponding residuals in the marginal distribution of residuals for group A. If υ_{Bi} is at the 70th percentile of the distribution of residuals of group B ($\tau_{Bi} = 0.7$), then $\upsilon_{Ai}^{C,2}$ will simply be the 70th percentile of the distribution of residuals for group A. In practice, most applications of the JMP procedure use this strong assumption of independence because there is little guidance on how a conditional imputation procedure could be used instead.

Limitations

Since independence of regression residuals is unrealistic, a more accurate implementation of JMP would require deciding how to condition on X when performing the

imputation procedure. If X consists of a limited number of groups or "cells", then one could perform the imputation within each of these groups. In general, however, it is difficult to know how to implement this ranking/imputation procedure in more general cases. As a result, other procedures such as the quantile method of Machado and Mata (2005) are increasingly being used as an alternative to JMP.

Another limitation of the JMP procedure is that there is no natural way of extending it to the case of the detailed decomposition for the composition effect.

Advantages

One advantage of the two-step procedure is that it provides a way of separating the between- and within-group components, as in a variance decomposition. This plays an important role in the inequality literature, since JMP concluded that most of the inequality growth from the 1960s to the 1980s was linked to the residual inequality component.

It is not clear, however, what is meant by between- and within-group components in the case of distributional measures like the 90-10 gap that are not decomposable. A better way of justifying JMP is that $Y = X\beta + \upsilon$ represents a structural model where X are observed skills, while υ represents unobserved skills. One can then perform simulation exercises asking what happens to the distribution when one either replaces returns to observed or unobserved skills (see also Section 2.2.3).

This economic interpretation also requires, however, some fairly strong assumptions. The two most important assumptions are the linearity of the model (Assumption 10, $m_g(X_i, \varepsilon_i) = X_i\beta_g + \upsilon_{gi}$) and rank preservation (Assumption 14). While linearity can be viewed as a useful approximation, rank preservation is much stronger since it means that someone with the same unobserved skills would be in the exact same position, conditional on X, in either group A or B. Just adding measurement error to the model would result in a violation of rank preservation.

Finally, if one is willing to interpret a simple regression as a decomposition between observed and unobserved skills, this can be combined with methods other than JMP. For instance, DFL perform regression adjustments to illustrate the effects of supply and demand factors on wages.[44]

4.4. Methods based on conditional quantiles

Procedure

Like JMP, Machado and Mata (2005, MM from hereinafter) propose a procedure based on transforming a wage observation Y_B into a counterfactual observation Y_A^C. The main advantage relative to JMP is that their estimation procedure, based on quantile regressions (Koenker and Bassett, 1978), provides an explicit way of estimating the (inverse) conditional distribution function $F^{-1}_{Y_A|X_A}(\cdot,\cdot)$ in the imputation function $g(Y,X) = F^{-1}_{Y_A|X_A}(F_{Y_B|X_B}(Y|X), X)$. One important difference, however, is that

[44] See also Lemieux (2002).

instead of transforming each actual observation of Y_{Bi} into a counterfactual Y_{Ai}^C, MM use a simulation approach where quantiles are drawn at random.

More specifically, since

$$Y_A^C = F_{Y_A|X_A}^{-1}(F_{Y_B|X_B}(Y|X), X),$$

and $\tau_B(Y|X) = F_{Y_B|X_B}(Y|X)$ follows a uniform distribution, one can think of doing the following:

1. Draw a simulated value τ_s from a uniform distribution $s = 1, \ldots, S$.
2. Estimate a linear quantile regression for the τ_sth quantile, and use the estimated result to predict simulated values of both Y_{Bs} and Y_{As}^C.[45] The reason for using quantile regressions is that:

$$Y_{As}^C = F_{Y_A|X_A}^{-1}(\tau_s, X) \quad \text{and} \quad Y_{Bs} = F_{Y_B|X_B}^{-1}(\tau_s, X),$$

   where $F_{Y_A|X_A}^{-1}(\cdot, \cdot)$ and $F_{Y_B|X_B}^{-1}(\cdot, \cdot)$ are the conditional quantile functions for the τ_sth quantile in group A and B, respectively.
3. Compare the simulated distributions of Y_{Bs} and Y_{As}^C to obtain measures of the wage structure effect. The composition effect is computed as the complement to the overall difference.

A key implementation question is how to specify the functional forms for the conditional quantile functions. MM suggest a linear specification in the X that can be estimated using quantile regression methods. The conditional quantile regression models can be written as:

$$Q_{g,\tau}(Y|X) = F_{Y_g|X_g}^{-1}(\tau, X) = X\beta_{g,\tau}, \quad g = A, B.$$

Table 4 reports in the top panel the results of the Machado-Mata procedure applied to our gender gap example using the male wage structure as reference.[46] It shows that the median gender log wage gap in the central column gives almost the same results for the aggregate decomposition as the OB mean gender log wage gap decomposition displayed in column (1) of Table 3. Going across the columns to compare quantile effects shows that gender differences in characteristics are much more important at the bottom (10th centile) than at the top (90th centile) of the wage distribution. Indeed, some significant wage structure effects emerge at the 90th percentile.

[45] For each random draw s, MM also draw a vector of covariates X_s from the observed data and perform the prediction for this value only. Melly (2005) discusses more efficient ways of computing distributions using this conditional quantile regression approach.

[46] The estimates were computed with Melly's implementation "rqdeco" in Stata.

Table 4 Gender wage gap: quantile decomposition results (NLSY, 2000).

Reference group: male coef.	10th percentile		50th percentile		90th percentile	
A: Raw log wage gap:						
$Q_\tau[\ln(w_m)] - Q_\tau[\ln(w_f)]$	0.170	(0.023)	0.249	(0.019)	0.258	(0.026)
B: Decomposition method: Machado-Mata-Melly						
Estimated log wage gap: $Q_\tau[\ln(w_m)] - Q_\tau[\ln(w_f)]$	0.192	(0.015)	0.239	(0.016)	0.276	(0.026)
Total explained by characteristics	0.257	(0.028)	0.198	(0.027)	0.143	(0.019)
Total wage structure	−0.065	(0.027)	0.041	(0.024)	0.133	(0.025)
C: Decomposition method: RIF **regressions without reweighting**						
Mean RIF gap: $E[\text{RIF}_\tau(\ln(w_m))] - E[\text{RIF}_\tau(\ln(w_f))]$	0.180	(0.023)	0.241	(0.019)	0.260	(0.026)
Composition effects attributable to						
Age, race, region, etc.	0.015	(0.005)	0.013	(0.004)	0.002	(0.004)
Education	−0.011	(0.005)	−0.017	(0.006)	−0.005	(0.010)
AFQT	0.005	(0.020)	0.013	(0.004)	0.013	(0.005)
L.T. withdrawal due to family	0.022	(0.021)	0.042	(0.014)	0.039	(0.017)
Life-time work experience	0.234	(0.026)	0.136	(0.014)	0.039	(0.023)
Industrial sectors	0.008	(0.012)	0.020	(0.008)	0.047	(0.011)
Total explained by characteristics	0.274	(0.035)	0.208	(0.025)	0.136	(0.028)
Wage structure effects attributable to						
Age, race, region, etc.	−0.342	(0.426)	0.168	(0.357)	0.860	(0.524)
Education	0.023	(0.028)	−0.030	(0.031)	0.023	(0.045)
AFQT	−0.007	(0.030)	0.003	(0.042)	0.008	(0.062)
L.T. withdrawal due to family	−0.075	(0.032)	−0.005	(0.025)	0.018	(0.032)
Life-time work experience	0.084	(0.148)	−0.085	(0.082)	−0.078	(0.119)
Industrial Sectors	0.015	(0.060)	−0.172	(0.046)	−0.054	(0.052)
Constant	0.208	(0.349)	0.154	(0.323)	−0.653	(0.493)
Total wage structure	−0.094	(0.044)	0.033	(0.028)	0.124	(0.036)

The data is an extract from the NLSY79 used in O'Neill and O'Neill (2006). Industrial sectors have been added to their analysis to illustrate issues linked to categorical variables. The other explanatory variables are age, dummies for black, hispanic, region, msa, central city. Bootstrapped standard errors are in parentheses. Means are reported in Table 2.

Limitations

This decomposition method is computationally demanding, and becomes quite cumbersome for data sets numbering more than a few thousand observations. Bootstrapping quantile regressions for sizeable number of quantiles τ (100 would be a minimum) is computationally tedious with large data sets. The implementation of the procedure can

be simplified by estimating a large number of quantile regressions (say 99, one for each percentile from 1 to 99) instead of drawing values of τ_s at random.[47]

Another limitation is that the linear specification is restrictive and finding the correct functional form for the conditional quantile regressions can be tedious. For instance, if there is a spike at the minimum wage in the wage distribution, this will result in flat spots in quantile regressions that would have to be captured with spline functions with knots that depend on X. Accurately describing a simple distribution with mass points (as is commonly observed in wage data) can, therefore, be quite difficult to do using quantile regressions.

As pointed out by Chernozhukov et al. (2009), it is not very natural to estimate *inverse* conditional distribution functions (quantile regressions) when the main goal of counterfactual exercises is to replace the conditional distribution function $F_{Y_B|X_B}$ with $F_{Y_A|X_A}$ to obtain Eq. (27). Chernozhukov et al. (2009) suggest instead to estimate directly distributional regression models for $F_{Y|X}(\cdot,\cdot)$, which is a more direct way of approaching the problem.

Advantages

One advantage of the MM approach is that it provides a natural way of performing a detailed decomposition for the wage structure component. The idea is to successively replace the elements of $\beta_{B,\tau}$ by those of $\beta_{A,\tau}$ when performing the simulations, keeping in mind that this type of detailed decomposition is path dependent. Unfortunately, the MM approach does not provide a way of performing the detailed decomposition for the composition effect.[48] This is a major drawback since the detailed decomposition of the composition effects is always clearly interpretable, while the detailed decomposition of the wage structure effect arbitrarily depends on the choice of the omitted group.

4.5. Reweighting methods

Procedure

As we mention in Section 4.2, another way of estimating the counterfactual distribution $F_{Y_A^C}(\cdot)$ is to replace the marginal distribution of X for group A with the marginal distribution of X for group B using a reweighting factor $\Psi(X)$. This idea was first introduced in the decomposition literature by DiNardo, Fortin and Lemieux [DFL] (1996). While DFL focus on the estimation of counterfactual densities in their empirical application, the method is easily applicable to any distributional statistic.

[47] See Melly (2005) for a detailed description of this alternative procedure. Gosling et al. (2000) and Autor et al. (2005) also use a similar idea in their empirical applications to changes in the distribution of wages over time.

[48] Machado and Mata (2005) (page 449-450) suggest computing the detailed decomposition for the composition effect using an unconditional reweighting procedure. This is invalid as a way of performing the decomposition for the same reason that a OB decomposition would be invalid if the β coefficient used for one covariate was estimated without controlling for the other covariates. We propose a conditional reweighting procedure in the next section that deals adequately with this issue.

In practice, the DFL reweighting method is similar to the propensity score reweighting method commonly used in the program evaluation literature (see Hirano et al. (2003)). For instance, in DFL's application to changes in wage inequality in the United States, time is viewed as a state variable, or in the context of the treatment effects literature as a treatment.[49] The impact of a particular factor or set of factors on changes in the wage distribution over time is constructed by considering the counterfactual state of the world where the distribution of this factor remained fixed in time, maintaining the Assumption 6 of invariance of the conditional distribution. Note that in contrast with the notation of this chapter, in DFL, time period 1 is used as reference group.[50] The choice of period 0 or period 1 as the reference group is analogous to the choice of whether the female or the male wage structure should be the reference wage structure in the analysis of the gender wage gap and is expected to yield different results in most cases.

In DFL, manipulations of the wage distributions, computed through reweighting, are applied to non-parametric estimates of the wage density, which can be particularly useful when local distortions, from minimum wage effects for example, are at play. To be consistent with the rest of this section, however, we focus our discussion on the cumulative distribution instead of the density. The key counterfactual distribution of interest, shown in Eq. (27) (distribution of wages that would prevail for workers in group A if they had the distribution of characteristics of group B) is constructed, as shown in Eq. (28), using the reweighting factor

$$\Psi(X) = \frac{\mathrm{d}F_{X_B}(X)}{\mathrm{d}F_{X_A}(X)}.$$

Although the reweighting factor is the ratio of two multivariate marginal distribution functions (of the covariates X), this expression can be simplified using Bayes' rule. Remembering that Bayes' rule states that

$$P(B_i|A) = P(A|B_i) \cdot P(B_i) / \sum_j P(A|B_j) \cdot P(B_j),$$

we have

$$\Pr(X|D_B = 1) = \frac{\Pr(D_B = 1|X) \cdot \mathrm{d}F(X)}{\int_x \Pr(D_B = 1|X) \cdot \mathrm{d}F(X)} = \frac{\Pr(D_B = 1|X)}{\Pr(D_B = 1)}$$

and a similar expression for $D_B = 0$. Since $\mathrm{d}F_{X_A}(X) = \Pr(X|D_B = 0)$ and $\mathrm{d}F_{X_B}(X) = \Pr(X|D_B = 1)$, the reweighting factor that keeps all conditioning variables as in

[49] This view of course makes more sense when some policy or other change has taken place over time (see Biewen, 2001).

[50] On the other hand, by analogy with the treatment effects literature, Firpo et al. (2007) use time period 0 as the reference group.

period 0 becomes

$$\Psi(X) = \frac{\Pr(X|D_B = 1)}{\Pr(X|D_B = 0)} = \frac{\Pr(D_B = 1|X)/\Pr(D_B = 1)}{\Pr(D_B = 0|X)/\Pr(D_B = 0)}.$$

The reweighting factor can be easily computed by estimating a probability model for $\Pr(D_B = 1|X)$, and using the predicted probabilities to compute a value $\widehat{\Psi}(X)$ for each observation. DFL suggest estimating a flexible logit or probit model, while Hirano, Imbens, and Ridder propose to use a non-parametric logit model.[51]

The reweighting decomposition procedure can be implemented in practice as follows:

1. Pool the data for group A and B and run a logit or probit model for the probability of belonging to group B:

$$\Pr(D_B = 1|X) = 1 - \Pr(D_B = 0|X) = 1 - \Pr(\varepsilon > -h(X)\beta) = \Lambda(-h(X)\alpha) \quad (30)$$

   where $\Lambda()$ is either a normal or logit link function, and $h(X)$ is a polynomial in X.
2. Estimate the reweighting factor $\widehat{\Psi}(X)$ for observations in group A using the predicted probability of belonging to group B ($\widehat{\Pr}(D_B = 1|X)$) and A ($\widehat{\Pr}(D_B = 0|X) = 1 - \widehat{\Pr}(D_B = 1|X)$), and the sample proportions in group B ($\widehat{\Pr}(D_B = 1)$) and A ($\widehat{\Pr}(D_B = 0)$):

$$\widehat{\Psi}(X) = \frac{\widehat{\Pr}(D_B = 1|X)/\widehat{\Pr}(D_B = 1)}{\widehat{\Pr}(D_B = 0|X)/\widehat{\Pr}(D_B = 0)}.$$

3. Compute the counterfactual statistic of interest using observations from the group A sample reweighted using $\widehat{\Psi}(X)$.

In DFL, the main object of interest is the probability density function, which is estimated using kernel density methods. The density for group A and the counterfactual density can be estimated as follows using kernel density methods, where $K(.)$ is the kernel function:[52]

$$\widehat{f}_{Y_A}(y) = \frac{1}{h \cdot N_A} \sum_{i \in A} K\left(\frac{Y_i - y}{h}\right),$$

$$\widehat{f}_{Y_A^C}(y) = \frac{1}{h \cdot N_A} \sum_{i \in A} \widehat{\Psi}(X_i) \cdot K\left(\frac{Y_i - y}{h}\right).$$

[51] The estimator suggested by Hirano et al. (2003) is a series estimator applied to the case of a logit model. The idea is to add increasingly higher order polynomial terms in the covariates as the size of the sample increases. Importantly, they also show that this approach yields an efficient estimate of the treatment effect.

[52] The two most popular kernel functions are the Gaussian and the Epanechnikov kernel.

Consider the density function for group A, $f_{Y_A}(y)$, and the counterfactual density $f_{Y_A^C}(y)$. The composition effect in a decomposition of densities is:

$$\Delta_X^{f(y)} = f_{Y_A^C}(y) - f_{Y_A}(y). \tag{31}$$

Various statistics from the wage distribution, such as the 10th, 50th, and 90th percentile, or the variance, Gini, or Theil coefficients can be computed either from the counterfactual density or the counterfactual distribution using the reweighting factor. The latter procedure is easier to use as it simply involves computing (weighted) statistics using standard computer packages. For example, the counterfactual variance can be computed as:

$$\widehat{\text{Var}}_{Y_A^C} = \frac{1}{N_A} \sum_{i \in A} \widehat{\Psi}(X_i) \cdot \left(Y_i - \widehat{\mu}_{Y_A^C}\right)^2,$$

where the counterfactual mean $\widehat{\mu}_{Y_A^C}$ is:

$$\widehat{\mu}_{Y_A^C} = \frac{1}{N_A} \sum_{i \in A} \widehat{\Psi}(X_i) \cdot Y_i.$$

For the 90-10, 90-50, and 50-10 wage differentials, the sought-after contributions to changes in inequality are computed as differences in the composition effects, for example,

$$\Delta_X^{90\text{-}10} = [Q_{A,.9}^C - Q_{A,.9}] - [Q_{A,.1}^C - Q_{A,.1}]. \tag{32}$$

Table 5 presents, in panel A, the results of a DFL decomposition of changes over time in male wage inequality using large samples from combined MORG-CPS data as in Firpo et al. (2007). In this decomposition, the counterfactual distribution of wages in 1983/85 is constructed by reweighting the characteristics of workers in 1983/85 (time period 0) so that they look like those of 2003/05 (time period 1) workers, holding the conditional distribution of wages in 1983/85 fixed.[53] The results of the aggregate decomposition, reported in the first three rows of Table 5, show that composition effects play a large role in changes in overall wage inequality, as measured by the 90-10 log wage differential or the variance of log wages. But the wage structure effects are more important when looking for increases at the top of the wage distribution, as measured by the 90-50 log wage differential, or decreases in the bottom, as measured by the 50-10 log wage differential.

[53] By contrast, in the original DiNardo et al. (1996) decomposition, workers in 1988 (time period 1) were reweighted to look like workers in 1979 (time period 0). The counterfactual distribution of wages was asking what would the distribution of wages look like if the workers' characteristics had remained at 1979 levels.

Table 5 Male wage inequality: aggregate decomposition results (CPS, 1983/85-2003/05)

Inequality measure	90-10		90-50		50-10		Variance		Gini	
A. Decomposition method: DFL - F(X) in 1983/85 reweighted to 2003/05										
Unadjusted change $(t_1 - t_0)$:	0.1091	(0.0046)	0.1827	(0.0037)	−0.0736	(0.0033)	0.0617	(0.0015)	0.0112	(0.0004)
Total composition effect	0.0756	(0.0031)	0.0191	(0.0034)	0.0565	(0.0029)	0.0208	(0.0007)	−0.0020	(0.0004)
Total wage effect	0.0336	(0.0048)	0.1637	(0.0043)	−0.1301	(0.0040)	0.0408	(0.0017)	0.0132	(0.0003)
B. Decomposition method: CFVM - LPM - $F^C(Y\|X) = \Lambda$ (2003/05 X's with 1983/85 α's)										
Estimated change $(t_1 - t_0)$:	0.1100	(0.0055)	0.1921	(0.0057)	−0.0821	(0.0044)	0.0636	(0.0013)	0.0118	(0.0005)
Total composition effect	0.0289	(0.0045)	0.0027	(0.0034)	0.0261	(0.0040)	0.0109	(0.0007)	−0.0046	(0.0003)
Total wage effect	0.0811	(0.0071)	0.1894	(0.0066)	−0.1082	(0.0060)	0.0527	(0.0016)	0.0164	(0.0006)
C. Decomposition method: CFVM - Logit - $F^C(Y\|X) = \Lambda$ (2003/05 X's with 1983/85 α's)										
Estimated change $(t_1 - t_0)$:	0.1100	(0.0040)	0.1921	(0.0032)	−0.0821	(0.0030)	0.0636	(0.0013)	0.0118	(0.0005)
Total composition effect	0.0872	(0.0044)	0.0392	(0.0040)	0.0480	(0.0018)	0.0212	(0.0008)	−0.0019	(0.0003)
Total wage effect	0.0227	(0.0053)	0.1529	(0.0049)	−0.1301	(0.0030)	0.0424	(0.0016)	0.0137	(0.0005)
D. Decomposition method: FFL- RIF-OLS - No reweighting										
Estimated change $(t_1 - t_0)$:	0.1100	(0.0039)	0.1824	(0.0036)	−0.0724	(0.0031)	0.0617	(0.0013)	0.0112	(0.0004)
Total composition effect	0.0617	(0.0018)	0.0294	(0.0019)	0.0323	(0.0014)	0.0151	(0.0005)	−0.0038	(0.0003)
Total wage effect	0.0483	(0.0043)	0.1530	(0.0043)	−0.1047	(0.0033)	0.0466	(0.0013)	0.0150	(0.0004)

The data is an extract from the Morg CPS 1983/85 (232 784 obs.) and 2003/05 (170 693 obs.) used in Firpo et al. (2007). The explanatory variables include union status, 6 education classes (high school omitted), 9 potential experience classes (20-25 years omitted). In Panel B and C, computations were performed using Melly's "counterfactual" procedure. The variance and Gini coefficient were computed using 100 quantile estimates. In Panel D, the estimated change is computed as $E[\text{RIF}_\tau(\ln(w_1))] - E[\text{RIF}_\tau(\ln(w_0))]$. Bootstrapped standard errors (100 reps.) are in parentheses.

Advantages

The main advantage of the reweighting approach is its simplicity. The aggregate decomposition for any distributional statistic is easily computed by running a single probability model (logit or probit) and using standard packages to compute distributional statistics with $\widehat{\Psi}(X_i)$ as weight.[54]

Another more methodological advantage is that formal results from Hirano et al. (2003) and Firpo (2007, 2010) establish the efficiency of this estimation method. Note that although it is possible to compute analytically the standard errors of the different elements of the decomposition obtained by reweighting, it is simpler in most cases to conduct inference by bootstrapping.[55]

For these two reasons, we recommend the reweighting approach as the method of choice for computing the aggregate decomposition. This recommendation even applies in the simple case of the mean decomposition. As pointed out by Barsky et al. (2002), a standard OB decomposition based on a linear regression model will yield biased estimates of the decomposition terms when the underlying conditional expectation of Y given X is non-linear (see Section 3.4). They suggest using a reweighting approach as an alternative, and the results of Hirano et al. (2003) can be used to show that the resulting decomposition is efficient.

Limitations

A first limitation of the reweighting method is that it is not straightforwardly extended to the case of the detailed decomposition. One exception is the case of binary covariates where it is relatively easily to compute the corresponding element of the decomposition. For instance, in the case of the union status (a binary covariate), DFL show how to compute the component of the composition corresponding to this particular covariate. It also relatively easy to compute the corresponding element of the wage structure effect. We discuss in Section 5 other options that can be used in the case of non-binary covariates.

As in the program evaluation literature, reweighting can have some undesirable properties in small samples when there is a problem of common support. The problem is that the estimated value of $\Psi(X)$ becomes very large when $\Pr(D_B = 1|X)$ gets close to 1. While lack of common support is a problem for any decomposition procedure, Frolich (2004) finds that reweighting estimators perform particularly poorly in this context, though Busso et al. (2009) reach the opposite conclusion using a different simulation experiment.[56]

[54] In small samples, it is important to ensure that these estimated weights sum up to the number of actual observations in the sample, though this is done automatically in packages like Stata. See Busso et al. (2009) for more detail.

[55] The analytical standard errors have to take account of the fact that the logit or probit model used to construct the reweighting factor is estimated. Firpo et al. (2007) show how to perform this adjustment. In practice, however, it is generally simpler to bootstrap the whole estimation procedure (both the estimation of the logit/probit to construct the weights and the computation of the various elements of the decomposition).

[56] In principle, other popular methods in the program evaluation literature such as matching could be used instead of reweighting.

Finally, even in cases where a pure reweighting approach has some limitations, there may be gains in combining reweighting with other approaches. For instance, we discuss in the next section how reweighting can be used to improve a decomposition based on the RIF-regression approach of Firpo et al. (2009). Lemieux (2002) also discusses how an hybrid approach based on DFL reweighting and the JMP decomposition procedure can be used to compute both the between- and within-group components of the composition and wage structure effects.

4.6. Methods based on estimating the conditional distribution

Procedure(s)

As mentioned above, when we first introduced the key counterfactual distribution of interest in Eq. (5), an alternative approach to the construction of this counterfactual is based on the estimation of the conditional distribution of the outcome variable, $F_{Y_A|X_A}(y|X)$. The counterfactual distribution is then estimated by integrating this conditional distribution over the distribution of X in group B.

Two early parametric methods based on this idea were suggested by Donald et al. (2000), and Fortin and Lemieux (1998).[57] Donald, Green and Paarsch propose estimating the conditional distribution using a hazard model. The (conditional) hazard function is defined as

$$h(y|X) = \frac{f(y|X)}{S(y|X)},$$

where $S(y|X) = 1 - F(y|X)$ is the survivor function. Therefore, the conditional distribution of the outcome variable, $F(y|X)$, or its density, $f(y|X)$, is easily recovered from the estimates of the hazard model. For instance, in the standard proportional hazard model[58]

$$h(y|X) = \exp(X\alpha)h_0(y),$$

estimates of α and of the baseline hazard $h_0(y)$ can be used to recover the conditional distribution

$$F(y|X) = 1 - \exp(-\Lambda_0(y)\exp(X\alpha)),$$

where $\Lambda_0(y) = \int h_0(u)\mathrm{d}u$ is the integrated baseline hazard.

[57] Foresi and Peracchi (1995) proposed to use a sequence of logit models to estimate the conditional distribution of excess returns.

[58] Donald et al. (2000) use a more general specification of the proportional hazard model where α and $h_0(y)$ are allowed to vary for different values (segments) of y.

Fortin and Lemieux (1998) suggest estimating an ordered probit model instead of a hazard model. They consider the following model for the outcome variable Y:

$$Y = \Lambda^{-1}(Y^*),$$

where $\Lambda(\cdot)$ is a monotonically increasing transformation function. The latent variable Y^*, interpreted as a latent "skill index" by Fortin and Lemieux, is defined as

$$Y^* = X\alpha + \varepsilon,$$

where ε is assumed to follow a standard normal distribution. It follows that the conditional distribution of Y is given by

$$F(y|X) = \Phi(-X\alpha + \Lambda(y)).$$

Fortin and Lemieux implement this in practice by discretizing the outcome variable into a large number of small bins. Each bin j corresponds to values of Y between the two thresholds c_{j-1} and c_j. The conditional probability of Y being in bin j is

$$\text{Prob}(c_{j-1} \leq Y \leq c_j|X) = \Phi(-X\alpha + \Lambda(c_j)) - \Phi(-X\alpha + \Lambda(c_{j-1})).$$

This corresponds to an ordered probit model where the $\Lambda(c_j)$ parameters (for $j = 1, \ldots, J$) are the usual latent variable thresholds. The estimated values of α and of the thresholds can then be used to construct the counterfactual distribution, just as in Donald et al. (2000).

To be more concrete, the following steps could be used to estimate the counterfactual distribution $F_{Y_A^C}(y)$ at the point $y = c_j$:

1. Estimate the ordered probit for group A. This yields estimates $\widehat{\alpha}_A$ and $\widehat{\Lambda}_A(c_j)$, the ordered probit parameters.
2. Compute the predicted probability $\widehat{F}_{Y_A|X_A}(c_j|X_i) = \Phi(-X_i\widehat{\alpha}_A + \widehat{\Lambda}_A(c_j))$ for each individual i in group B.
3. For each threshold c_j, compute the sample average of $\widehat{F}_{Y_A|X_A}(c_j|X_i)$ over all observations in group B:

$$\widehat{F}_{Y_A^C}(c_j) = \frac{1}{N_B}\sum_{i\in B}\Phi(-X_i\widehat{\alpha}_A + \widehat{\Lambda}_A(c_j)).$$

Repeating this for a large number of values of $y = c_j$ will provide an estimate of the counterfactual distribution $F_{Y_A^C}(y)$.

In a similar spirit, Chernozhukov et al. (2009) suggest a more flexible distribution regression approach for estimating the conditional distribution $F(y|X)$. Following

Foresi and Peracchi (1995), the idea is to estimate a separate regression model for each value of y. They consider the model $F(y|X) = \Lambda(X\alpha(y))$, where $\Lambda(\cdot)$ is a known link function. For example, if $\Lambda(\cdot)$ is a logistic function, $\alpha(y)$ can be estimated by creating a dummy variable $\mathbb{1}\{Y_i \leq y\}$ indicating whether the value of Y_i is below y, where $\mathbb{1}\{\cdot\}$ is the indicator function, and running a logit regression of $\mathbb{1}\{Y_i \leq y\}$ on X_i to estimate $\alpha(y)$.

Similarly, if the link function is the identity function ($\Lambda(z) = z$) the probability model is a linear probability model. If the link function is the normal CDF ($\Lambda(z) = \Phi(z)$) the probability model is a probit. Compared to Fortin and Lemieux (1998), Chernozhukov et al. (2009) suggest estimating a separate probit for each value of y, while Fortin and Lemieux use a more restrictive model where only the intercept (the threshold in the ordered probit) is allowed to change for different values of y.

As above, the counterfactual distribution can be obtained by first estimating the regression model (probit, logit, or LPM) for group A to obtain the parameter estimates $\widehat{\alpha}_A(y)$, computing the predicted probabilities $\Lambda(X_i\widehat{\alpha}_A(y))$, and averaging over these predicted probabilities to get the counterfactual distribution $\widehat{F}_{Y_A^C}(y)$:

$$\widehat{F}_{Y_A^C}(y) = \frac{1}{N_B}\sum_{i\in B}\Lambda(X_i\widehat{\alpha}_A(y)).$$

Once the counterfactual distribution $\widehat{F}_{Y_A^C}(y)$ has been estimated, counterfactual quantiles can be obtained by inverting the estimated distribution function. Consider $Q^C_{\tau,A}$, the τth quantile of the counterfactual distribution $F_{Y_A^C}(\cdot)$. The estimated counterfactual quantile is:

$$\widehat{Q}^C_{A,\tau} = \widehat{F}^{-1}_{Y_A^C}(\tau).$$

It is useful to illustrate graphically how the estimation of the counterfactual distribution $\widehat{F}_{Y_A^C}(y)$ and the inversion into quantiles can be performed in practice. Figure 1 first shows the actual CDF's for group A, $F_{Y_A}(\cdot)$, and B, $F_{Y_B}(\cdot)$, respectively. The squares in between the two cumulative distributions illustrate examples of counterfactuals computed using the one of the method discussed above.

For example, consider the case of the median wage for group B, $Q_{B,.5}$. Using the distribution regression approach of Chernozhukov et al. (2009), one can estimate, for example, a LPM by running a regression of $\mathbb{1}\{Y_i \leq Q_{B,.5}\}$ on X_i for group A. This yields an estimate of $\widehat{\alpha}_A(y = Q_{B,.5})$ that can then be used to compute $\widehat{F}_{Y_A^C}(y = Q_{B,.5})$. This counterfactual proportion is represented by the square on the vertical line over $y = Q_{B,.5}$ in Fig. 1.

Figure 2 then illustrates what happens when a similar exercise is performed for a larger number of values of y (100 in this particular figure). It now becomes clear from

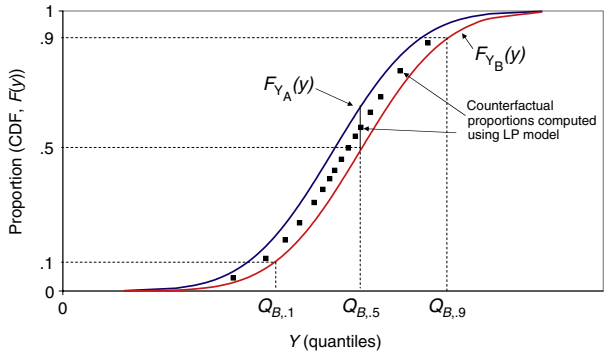


Figure 1 ***Relationship between proportions and quantiles.***

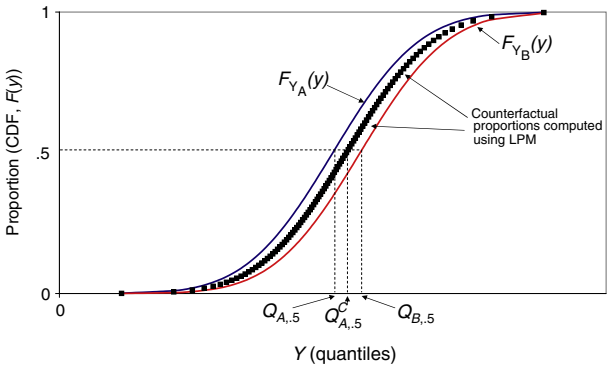


Figure 2 ***Inverting globally.***

the figure how to numerically perform the inversion. In the case of the median, the total gap between group A and B is $Q_{B,.5} - Q_{A,.5}$. The counterfactual median can then be estimated by picking the corresponding point $Q^C_{A,.5}$ on the counterfactual function defined by the set of points estimated by running a set of LPM at different values of y. In practice, one could compute the precise value of $Q^C_{A,.5}$ by estimating the LPMs (or a logit or probit) for a large number of values of y, and then "connecting the dots" (i.e. using linear interpolations) between these different values.

Figure 2 also illustrates one of the key messages of the chapter listed in the introduction, namely that is it easier to estimate models for proportions than quantiles. In Fig. 2, the difference in the proportion of observations under a given value of y is simply the vertical distance between the two cumulative distributions, $F_{Y_B}(y) - F_{Y_A}(y)$. Decomposing this particular gap in proportion is not a very difficult problem. As discussed in Section 3.5, one can simply run a LPM and perform a standard OB decomposition. An alternative also discussed in Section 3.5 is to perform a nonlinear decomposition

using a logit or probit model. The conditional distribution methods of Fortin and Lemieux (1998) and Chernozhukov et al. (2009) essentially amount to computing this decomposition in the vertical dimension.

By contrast, it is not clear at first glance how to decompose the *horizontal* distance, or quantile gap, between the two curves. But since the vertical and horizontal are just two different ways of describing the same difference between the two cumulative distributions $F_{Y_B}(y)$ and $F_{Y_A}(y)$, one can perform a first decomposition either vertically or horizontally, and then invert back to get the decomposition in the other dimension. Since decomposing proportions (the vertical distance) is relatively easy, this suggests first performing the decomposition on proportions at many points of the distribution, and then inverting back to get the decomposition in the quantile dimension (the horizontal distance).

Table 5 reports, in panels B and C, the results of the aggregate decomposition results for male wages using the method of Chernozhukov et al. (2009). The counterfactual wage distribution is constructed by asking what would be the distribution of wages in 1983/85 if the conditional distribution was as in 2003/05. Panel B uses the LPM to estimate $\Lambda(X_i\widehat{\alpha}_A(y))$ while the logit model is used in Panel C.[59] The first rows of Panel B and C show the changes in the wage differentials based on the fitted distributions, so that any discrepancies between these rows in the first row of Panel A shows the estimation errors. The second rows report the composition effects computed as the difference between the fitted distribution in 1983/85 and the counterfactual distribution. Given our relatively large sample, the differences across estimators in the different panels are at times statistically different. However, the results from the logit estimation in Panel C give results that are qualitatively similar to the DFL results shown in Panel A, with composition effects being relatively more important in accounting for overall wage inequality, as measured by the 90-10 log wage differential, and wage structure effects playing a relatively more important role in increasing wage inequality at the top and reducing wage inequality at the bottom.

Limitations

If one is just interested in performing an aggregate distribution, it is preferable to simply use the reweighting methods discussed above. Like the conditional quantile methods discussed in Section 4.4, conditional distribution methods require some parametric assumptions on the distribution regressions that may or may not be valid. Chernozhukov, Fernandez-Val, and Melly's distribution regression approach is more flexible than earlier suggestions by Donald et al. (2000) and Fortin and Lemieux (1998), but it potentially involves estimating a large number of regressions.

[59] The estimation was performed using Melly's "counterfactual" Stata procedure. The computation of the variance and Gini coefficient were based on the estimation of 100 centiles.

Running unconstrained regressions for a large number of values of y may result, however, in non-monotonicities in the estimated counterfactual distribution $\widehat{F}_{Y_A^C}(y)$. Smoothing or related methods then have to be used to make sure that the counterfactual distribution is monotonic and, thus, invertible into quantiles.[60] By contrast, reweighting methods require estimating just one flexible logit or probit regression, which is very easy to implement in practice.

Advantages

An important advantage of distribution regression methods over reweighting is that they can be readily generalized to the case of the detailed decomposition, although these decomposition will be path dependent. We show in the next section how Chernozhukov, Fernandez-Val, and Melly's distribution regression approach, and the related RIF regression method of Firpo et al. (2009) can be used to perform a detailed decomposition very much in the spirit of the traditional OB decomposition for the mean.

4.7. Summary

In this section we discuss most of the existing methods that have been proposed to perform an aggregate decomposition for general distributional statistics. While all these methods could, in principle, yield similar results, we argue that DFL reweighting is the method of choice in this context for two main reasons. First, it is simple to implement as it simply involves estimating a single logit or probit model for computing the reweighting factors. Counterfactual values of any distributional statistical can then be readily computed from the reweighted sample. By contrast, methods that yield counterfactual estimates of quantiles or the whole CDF require estimating a separate model at a large number of points in the distribution.

The second advantage of reweighting is that there are well established results in the program evaluation that show that the method is asymptotically efficient (Hirano et al., 2003; Firpo, 2007).

5. DETAILED DECOMPOSITIONS FOR GENERAL DISTRIBUTIONAL STATISTICS

In this section, we extend the methods introduced above for the aggregate decomposition to the case of the detailed decomposition. We first show that conditional distribution methods based on distribution regressions can be used to compute both the composition and wage structure subcomponents of the detailed decomposition. We then discuss a related method based the RIF-regressions introduced in Firpo et al. (2009). The main advantage of this last procedure is that it is regression based and, thus, as easy to use in practice as the traditional OB method.

[60] Chernozhukov et al. (2009) use the method of Chernozhukov et al. (2010) to ensure that the function is monotonic.

The other methods proposed in Section 4 are not as easy to extend to the case of the detailed decomposition. We discuss, nonetheless, which elements of the detailed decomposition can be estimated using these various methods, and under which circumstances it is advantageous to use these methods instead of others.

5.1. Methods based on the conditional distribution

Procedure

In the case where the specification used for the distribution regression is the LPM, the aggregate decomposition of Section 4.6 can be generalized to the detailed decomposition as follows. Since the link function for the LPM is $\Lambda(z) = z$, the counterfactual distribution used earlier becomes:

$$\widehat{F}_{Y_A^C}(y) = \frac{1}{N_B} \sum_{i \in B} X_i \widehat{\alpha}_A(y) = \overline{X}_B \widehat{\alpha}_A(y).$$

We can also write:

$$\begin{aligned} \widehat{F}_{Y_B}(y) - \widehat{F}_{Y_A}(y) &= \left[\widehat{F}_{Y_B}(y) - \widehat{F}_{Y_A^C}(y)\right] + \left[\widehat{F}_{Y_A^C}(y) - \widehat{F}_{Y_A}(y)\right] \\ &= \overline{X}_B \left(\widehat{\alpha}_B(y) - \widehat{\alpha}_A(y)\right) + \left(\overline{X}_B - \overline{X}_A\right) \widehat{\alpha}_A(y), \end{aligned}$$

where the first term is the familiar wage structure effect, while the second term is the composition effect. The above equation can, therefore, be used to compute a detailed decomposition of the difference in the proportion of workers below wage y between groups A and B. We obtain the detailed distribution of quantiles by (i) computing the different counterfactuals for each element of X and α sequentially, for a large number of values of y, and (ii) inverting to get the corresponding quantiles for each detailed counterfactual. A similar approach could also be used when the link function is a probit or a logit by using the procedure suggested in Section 3.5.

Advantages

The main advantage of this method based on distribution regressions and the global inversion of counterfactual CDF into counterfactual quantiles (as in Fig. 2) is that it yields a detailed decomposition comparable to the OB decomposition of the mean.

Limitations

One limitation of this method is that it involves computing a large number of counterfactuals CDFs and quantiles, as the procedure has to be repeated for a sizable number of values of y. This can become cumbersome because of the potential non-monotonicity problems discussed earlier. Furthermore, the procedure suffers from the problem of path dependence since the different counterfactual elements of the detailed decomposition

have to be computed sequentially. For these reasons, we next turn to a simpler approach based on a local, as opposed to a global, inversion of the CDF.

5.2. RIF-regression methods

Procedure

RIF-regression methods provide a simple way of performing detailed decompositions for any distributional statistic for which an influence function can be computed. Although we focus below on the case of quantiles of the unconditional distribution of the outcome variable, our empirical example includes the case of the variance and Gini. The procedure can be readily used to address glass ceiling issues in the context of the gender wage gap, or changes in the interquartile range in the context of changes in wage inequality. It can be used to either perform OB- type detailed decompositions, or a slightly modified "hybrid" version of the decomposition suggested by Firpo et al. (2007) (reweighting combined with RIF regressions, as in Section 3.4 for the mean).

A RIF-regression (Firpo et al., 2009) is similar to a standard regression, except that the dependent variable, Y, is replaced by the (recentered) influence function of the statistic of interest. Consider $\text{IF}(y;\nu)$, the influence function corresponding to an observed wage y for the distributional statistic of interest, $\nu(F_Y)$. The recentered influence function (RIF) is defined as $\text{RIF}(y;\nu) = \nu(F_Y) + \text{IF}(y;\nu)$, so that it aggregates back to the statistics of interest ($\int \text{RIF}(y;\nu)\cdot \text{d}F(y) = \nu(F_Y)$). In its simplest form, the approach assumes that the conditional expectation of the $\text{RIF}(Y;\nu)$ can be modeled as a linear function of the explanatory variables,

$$\mathbb{E}\left[\text{RIF}(Y;\nu)\,|X\right] = X\gamma,$$

where the parameters γ can be estimated by OLS.[61]

In the case of quantiles, the influence function $\text{IF}(Y, Q_\tau)$ is given by $(\tau - \mathbb{1}\{Y \leq Q_\tau\})/f_Y(Q_\tau)$, where $\mathbb{1}\{\cdot\}$ is an indicator function, $f_Y(\cdot)$ is the density of the marginal distribution of Y, and Q_τ is the population τ-quantile of the unconditional distribution of Y. As a result, $\text{RIF}(Y;Q_\tau)$ is equal to $Q_\tau + \text{IF}(Y, Q_\tau)$, and can be rewritten as

$$\text{RIF}(y;Q_\tau) = Q_\tau + \frac{\tau - \mathbb{1}\{y \leq Q_\tau\}}{f_Y(Q_\tau)} = c_{1,\tau}\cdot\mathbb{1}\{y > Q_\tau\} + c_{2,\tau}, \tag{33}$$

where $c_{1,\tau} = 1/f_Y(Q_\tau)$ and $c_{2,\tau} = Q_\tau - c_{1,\tau}\cdot(1-\tau)$. Except for the constants $c_{1,\tau}$ and $c_{2,\tau}$, the RIF for a quantile is simply an indicator variable $\mathbb{1}\{Y \leq Q_\tau\}$ for whether the outcome variable is smaller or equal to the quantile Q_τ. Using the terminology introduced above, running a linear regression of $\mathbb{1}\{Y \leq Q_\tau\}$ on X is a distributional regression estimated at $y = Q_\tau$, using the link function of the linear probability model ($\Lambda(z) = z$).

[61] Firpo et al. (2009) also propose other more flexible estimation procedures.

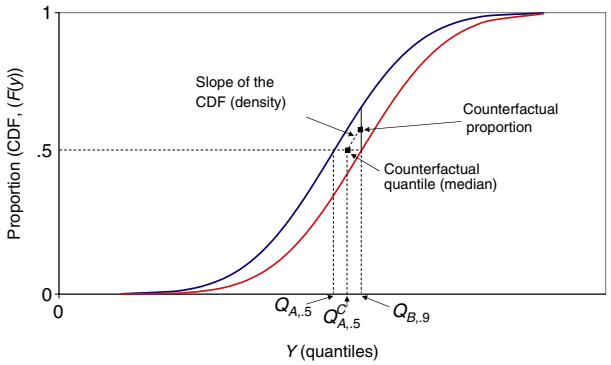


Figure 3 RIF ***regressions: Inverting locally.***

There is, thus, a close connection between RIF regressions and the distributional regression approach of Chernozhukov et al. (2009). In both cases, regression models are estimated for explaining the determinants of the proportion of workers earning less than a certain wage. As we saw in Fig. 2, in Chernozhukov et al. (2009) estimates of models for proportions are then *globally* inverted back into the space of quantiles. This provides a way of decomposing quantiles using a series of simple regression models for proportions.

Figure 3 shows that RIF-regressions for quantiles are based on a similar idea, except that the inversion is only performed *locally*. Suppose that after estimating a model for proportions, we compute a counterfactual proportion based on changing either the mean value of a covariate, or the return to the covariate estimated with the LPM regression. Under the assumption that the relationship between counterfactual proportions and counterfactual quantiles is locally linear, one can then go from the counterfactual proportion to the counterfactual quantile (both illustrated in Fig. 3) by moving along a line with a slope given by the slope of the counterfactual distribution function. Since the slope of a cumulative distribution is the just the probability density function, one can easily go from proportions to quantiles by dividing the elements of the decomposition for proportions by the density.

While the argument presented in Fig. 3 is a bit heuristic, it provides the basic intuition for how we can get a decomposition model for quantiles by simply dividing a model for proportions by the density. As we see in Eq. (33), in the RIF for quantiles, the indicator variable $\mathbb{1}\{y \leq Q_\tau\}$ is indeed divided by $f_Y(Q_\tau)$ (i.e. multiplying by the constant $c_{1,\tau}$).

Firpo et al. (2009) explain how to first compute the RIF, and then run regressions of the RIF on the vector of covariates. In the case of quantiles, the RIF is first estimated by computing the sample quantile $\widehat{Q}_\tau$, and estimating the density at that point using kernel methods. An estimate of the RIF of each observation, $\widehat{\text{RIF}}(Y_i; Q_\tau)$, is then obtained by plugging the estimates $\widehat{Q}_\tau$ and $\widehat{f}(\widehat{Q}_\tau)$ into Eq. (33).

Letting the coefficients of the unconditional quantile regressions for each group be

$$\widehat{\gamma}_{g,\tau} = \left(\sum_{i\in G} X_i \cdot X_i^{\intercal}\right)^{-1} \cdot \sum_{i\in G} \widehat{\mathrm{RIF}}(Y_{gi}; Q_{g,\tau}) \cdot X_i, \quad g = A, B \tag{34}$$

we can write the equivalent of the OB decomposition for any unconditional quantile as

$$\widehat{\Delta}_O^{\tau} = \overline{X}_B \left(\widehat{\gamma}_{B,\tau} - \widehat{\gamma}_{A,\tau}\right) + \left(\overline{X}_B - \overline{X}_A\right) \widehat{\gamma}_{A,\tau} \tag{35}$$

$$= \widehat{\Delta}_S^{\tau} + \widehat{\Delta}_X^{\tau}. \tag{36}$$

The second term in Eq. (36) can be rewritten in terms of the sum of the contribution of each covariate as

$$\widehat{\Delta}_X^{\tau} = \sum_{k=1}^{K} \left(\overline{X}_{Bk} - \overline{X}_{Ak}\right) \widehat{\gamma}_{Ak,\tau}.$$

That is, the detailed elements of the composition effect can be computed in the same way as for the mean. Similarly, the detailed elements of the wage structure effects can be computed, but as in the case of the mean, these will also be subject to the problem of the omitted group.

Table 4 presents in its bottom panel such OB like gender wage gap decomposition of the 10th, 50th, and 90th percentiles of the unconditional distribution of wages corresponding to Tables 2 and 3 using the male coefficients as reference group and without reweighting. As with the MM decomposition presented in the top panel, the composition effects from the decomposition of the median gender pay gap reported in the central column of Table 4 are very close to those of the decomposition of the mean gender pay gap reported in column (1) of Table 3. As before, the wage structure effects in the relatively small NLSY sample are generally not statistically significant, with the exception of the industrial sectors which are, however, subject to the categorical variables problem. The comparison of the composition effects at the 10th and 90th percentiles shows that the impact of differences in life-time work experience is much larger at the bottom of the distribution than at the top where it is not statistically significant. Note that the aggregate decomposition results obtained using either the MM method or the RIF regressions do not exhibit statistically significant differences. Table 5 presents in Panel D the results of the aggregate decomposition using RIF-regressions without reweighting. The results are qualitatively similar to those of Panels A and C. Table 6 extends the analysis of the decomposition of male wage inequality presented in Table 5 to the detailed decomposition. For each inequality measures, the detailed decomposition are presented both for the extension of the classic OB decomposition in

Table 6 Male wage inequality: FFL decomposition results (CPS, 1983/85-2003/05).

Inequality measure	**90-10**				**90-50**				**50-10**			
Reweighting	*No reweighting 1983/85 reference*		*$F(X)$ in 1983/85 reweighted to 2003/05*		*No reweighting 1983/85 reference*		*$F(X)$ in 1983/85 reweighted to 2003/05*		*No reweighting 1983/85 reference*		*$F(X)$ in 1983/85 reweighted to 2003/05*	
Unadjusted change	0.1100	(0.0039)	0.1100	(0.0039)	0.1824	(0.0036)	0.1824	(0.0036)	−0.0724	(0.0031)	−0.0724	(0.0031)
Composition effects attributable to												
Union	0.0353	(0.0008)	0.0356	(0.0008)	0.0568	(0.0009)	0.0572	(0.0009)	−0.0215	(0.0006)	−0.0216	(0.0006)
Education	0.0150	(0.0012)	0.0149	(0.0012)	−0.0155	(0.0010)	−0.0156	(0.0010)	0.0305	(0.0010)	0.0305	(0.0010)
Experience	0.0114	(0.0009)	0.0114	(0.0009)	0.0119	(0.0009)	0.0118	(0.0009)	0.0233	(0.0008)	0.0233	(0.0008)
Total explained	0.0617	(0.0018)	0.0619	(0.0018)	0.0294	(0.0019)	0.0298	(0.0019)	0.0323	(0.0014)	0.0322	(0.0013)
Wage structure effects attributable to												
Union	0.0019	(0.0016)	0.0084	(0.0016)	0.0016	(0.0018)	0.0141	(0.0018)	0.0035	(0.0014)	0.0225	(0.0014)
Education	0.1053	(0.0068)	0.1234	(0.0064)	0.0339	(0.0070)	0.0754	(0.0067)	0.0714	(0.0053)	0.0480	(0.0059)
Experience	0.0115	(0.0127)	−0.0768	(0.0138)	−0.0120	(0.0110)	−0.0451	(0.0116)	0.0235	(0.0081)	0.0318	(0.0092)
Constant	−0.0705	(0.0148)	−0.0211	(0.0158)	0.1326	(0.0129)	0.1477	(0.0134)	−0.2031	(0.0095)	−0.1688	(0.0113)
Total wage structure	0.0483	(0.0043)	0.0339	(0.0042)	0.1530	(0.0043)	0.1639	(0.0043)	−0.1047	(0.0033)	−0.1300	(0.0039)

(*continued on next page*)

Table 6 (continued)

Inequality measure	**Variance**				**Gini**			
Reweighting	*No reweighting 1983/85 reference*		*$F(X)$ in 1983/85 reweighted to 2003/05*		*No reweighting 1983/85 reference*		*$F(X)$ in 1983/85 reweighted to 2003/05*	
Unadjusted change	0.0617	(0.0013)	0.0617	(0.0013)	0.0112	(0.0004)	0.0112	(0.0004)
Composition effects attributable to								
Union	0.0129	(0.0002)	0.0130	(0.0002)	0.0069	(0.0001)	0.0069	(0.0001)
Education	0.0013	(0.0003)	0.0013	(0.0003)	−0.0058	(0.0001)	−0.0058	(0.0001)
Experience	0.0009	(0.0003)	0.0009	(0.0003)	−0.0049	(0.0001)	−0.0049	(0.0001)
Total explained	0.0151	(0.0005)	0.0152	(0.0005)	−0.0038	(0.0003)	−0.0037	(0.0003)
Wage structure effects attributable to								
Union	0.0002	(0.0005)	0.0023	(0.0005)	0.0020	(0.0001)	0.0011	(0.0001)
Education	0.0483	(0.0020)	0.0419	(0.0020)	0.0070	(0.0007)	0.0064	(0.0007)
Experience	0.0033	(0.0041)	−0.0177	(0.0041)	−0.0003	(0.0011)	−0.0064	(0.0012)
Constant	−0.0052	(0.0048)	0.0145	(0.0048)	0.0063	(0.0014)	0.0129	(0.0014)
Total wage structure	0.0466	(0.0013)	0.041	(0.0013)	0.0150	(0.0004)	0.0132	(0.0004)

The data is an extract from the Morg CPS 1983/85 (232 784 obs.) and 2003/05 (170 693 obs.) used in Firpo et al. (2007). The explanatory variables include union status, 6 education classes (high school omitted), 9 potential experience classes (20-25 years omitted). Bootstrapped standard errors (100 reps.) are in parentheses.

Eq. (36), and for the reweighted-regression decomposition, described in the case of the mean in Section 3.4.[62] For the reweighted-regression decomposition, Table 6 reports the detailed elements of the main composition effect $\widehat{\Delta}^{\tau}_{X,p}$ and the detailed elements of the main wage structure effect $\widehat{\Delta}^{\tau}_{S,p}$, where

$$\widehat{\Delta}^{\tau}_{X,p} = \left(\overline{X}^{C}_{A} - \overline{X}_{A}\right)\widehat{\gamma}_{A,\tau} \quad \text{and} \quad \widehat{\Delta}^{\tau}_{S,p} = \overline{X}_{B}\left(\widehat{\gamma}_{B,\tau} - \widehat{\gamma}^{C}_{A,\tau}\right),$$

and where the group A sample is reweighted to mimic the group B sample, which means we should have $plim(\overline{X}^{C}_{A}) = plim(\overline{X}_{B})$. The total reweighting error $\widehat{\Delta}^{\tau}_{S,e} = \left(\overline{X}_{B} - \overline{X}^{C}_{A}\right)\widehat{\gamma}^{C}_{A,\tau}$ corresponds to the difference between the "Total explained" across the classic OB and the reweighted-regression decomposition. For example, for the 90-10 log wage differential, it is equal to $0.0617 - 0.0619 = 0.0002$.[63] The total specification error, $\widehat{\Delta}^{\tau}_{X,e} = \overline{X}^{C}_{A}(\widehat{\gamma}^{C}_{A,\tau} - \widehat{\gamma}_{A,\tau})$, corresponds to the difference between the "Total wage structure" across the classic OB and the reweighted-regression decomposition and is found to be more important. In terms of composition effects, de-unionization is found to be an important factor accounting for the polarization of male wage inequality. It is also found to reduce inequality at the bottom, as measured by the 50-10 log wage differential, and to increase inequality at the top, as measured by the 90-50 log wage differential. In terms of wage structure effects, increases in the returns to education are found, as in Lemieux (2006a), to be the dominant factor accounting for overall increases in male wage inequality.

Advantages

The linearity of RIF regressions has several advantages. It is straightforward to invert the proportion of interest by dividing by the density. Since the inversion can be performed locally, another advantage is that we don't need to evaluate the global impact at all points of the distribution and worry about monotonicity. One gets a simple regression which is easy to interpret. As a result, the resulting decomposition is path independent.

Limitations

Like many other methods, RIF regressions assume the invariance of the conditional distribution (i.e., no general equilibrium effects). Also, a legitimate practical issue is how good the approximation is. For relatively smooth dependent variables, such as test scores, it may be a moot point. But in the presence of considerable heaping (usually displayed

[62] Using a reweighted regression approach can be particularly important in the cases of RIF-regressions that are unlikely to be linear for distributional statistics besides the mean.

[63] The reweighting error reflects the fact that the composition effect in the reweighted-regression decomposition, $\left(\overline{X}_{B} - \overline{X}^{C}_{A}\right)\widehat{\gamma}^{C}_{A,\tau}$, is not exactly equal to the standard composition effect $\left(\overline{X}_{B} - \overline{X}_{B}\right)\widehat{\gamma}^{C}_{A,\tau}$ when the reweighted mean $\overline{X}^{C}_{A}$ is not exactly equal to $\overline{X}_{B}$.

in wage distribution), it may advisable to oversmooth density estimates and compare its values around the quantile of interest. This can be formally looked at by comparing reweighting estimates to the OB-type composition effect based on RIF regressions (the specification error discussed earlier).

5.3. A reweighting approach

Procedure(s)

As we mention in Section 4, it is relatively straightforward to extend the DFL reweighting method to perform a detailed decomposition in the case of binary covariates. DFL show how to compute the composition effect corresponding to a binary covariate (union status in their application). Likewise, DiNardo and Lemieux (1997) use yet another reweighting technique to compute the wage structure component. We first discuss the case where a covariate is a binary variable, and then discuss the case of categorical (with more than 2 categories) and continuous variables.

Binary covariate

Consider the case of one binary covariate, X_1, and a vector of other covariates, X_2. For instance, DiNardo et al. (1996) look at the case of unionization. They are interested in isolating the contribution of de-unionization to the composition effect by estimating what would have happened to the wage distribution if the distribution of unionization, but of none of the other covariates, had changed over time.

Letting A index the base period and B the end period, consider the counterfactual distribution $F_{Y_A^{C,X_1}}$, which represents the period A distribution that would prevail if the conditional distribution of unionization (but of none of the other covariates X_2) was as in period B.[64] Note that we are performing a counterfactual experiment by changing the conditional, as opposed to the marginal, distribution of unionization. Unless unionization is independent of other covariates ($X_1 \perp X_2$), the marginal distribution of unionization, $F_X(X_1)$, will depend on the distribution of X_2, $F_X(X_2)$. For instance, if unionization is higher in the manufacturing sector, but the share of workers in manufacturing declines over time, the overall unionization rate will decline even if, conditional on industrial composition, the unionization rate remains the same.

Using the language of program evaluation, we want to make sure that secular changes in the rate of unionization are not confounded by other factors such as industrial change. This is achieved by looking at changes in the conditional, as opposed to the marginal distribution of unionization. Note that the main problem with the procedure suggested by MM to compute the elements of the composition effect corresponding to each covariate is that it fails to control this problem. MM suggest using an unconditional reweighting procedure based on the change in the marginal, as opposed to the

[64] Note that in DFL, it is the opposite; group B is the 1988 time period and group A is the 1979 time period.

conditional distribution of covariates. Unless the covariates are independent, this will yield biased estimates of the composition effect elements of the detailed decomposition.

The counterfactual distribution $F_{Y_A^{C,X_1}}$ is formally defined as

$$\begin{aligned} F_{Y_A^{C,X_1}}(y) &= \int\int F_{Y_A|X_A}(y|X_1, X_2)\, \mathrm{d}F_{X_B}(X_1|X_2)\, \mathrm{d}F_{X_A}(X_2) \\ &= \int\int F_{Y_A|X_A}(y|X_1, X_2)\ \Psi_1(X_1, X_2)\, \mathrm{d}F_{X_A}(X_1|X_2)\, \mathrm{d}F_{X_A}(X_2) \\ &= \int\int F_{Y_A|X_A}(y|X_1, X_2)\ \Psi_1(X_1, X_2)\, \mathrm{d}F_{X_A}(X_1, X_2), \end{aligned}$$

where the reweighting function is

$$\Psi_{X_1}(X_1, X_2) \equiv \frac{\mathrm{d}F_{X_B}(X_1|X_2)}{\mathrm{d}F_{X_A}(X_1|X_2)} \tag{37}$$

$$= X_1 \cdot \frac{\Pr_B(X_1 = 1|X_2)}{\Pr_A(X_1 = 1|X_2)} + (1 - X_1) \cdot \frac{\Pr_B(X_1 = 0|X_2)}{\Pr_A(X_1 = 0|X_2)}. \tag{38}$$

Note that the conditional distribution $F_{Y_A}(Y_A|X_1, X_2)$ is assumed to be unaffected by the change in the conditional distribution of unionization (assumption of invariance of conditional distribution in Section 2). This amounts to assuming away selection into union status based on unobservables (after controlling for the other covariates X_2).

The reweighting factor $\Psi_1(X_1, X_2)$ can be computed in practice by estimating two probit or logit models for the probability that a worker is unionized in period A and B, respectively. The resulting estimates can then be used to compute the predicted probability of being unionized ($\Pr_A(X_1 = 1|X_2)$ and $\Pr_B(X_1 = 1|X_2)$) or not unionized ($\Pr_A(X_1 = 0|X_2)$ and $\Pr_B[X_1 = 0|X_2]$), and then plugging these estimates into the above formula.

DiNardo and Lemieux (1997) use a closely related reweighting procedure to compute the wage structure component of the effect of unions on the wage distribution. Consider the question of what would happen to the wage distribution if no workers were unionized. The distribution of wages among non-union workers:

$$F_{Y_g}(y|X_1 = 0) = \int\int F_{Y_g|X_g}(y|X_1 = 0, X_2)\, \mathrm{d}F_{X_A}(X_2|X_1 = 0),$$

is not a proper counterfactual since the distribution of other covariates, X_2, may not be the same for union and non-union workers. DiNardo and Lemieux (1997) suggest solving this problem by reweighting non-union workers so that their distribution of X_2 is the same as for the entire workforce. The reweighting factor that accomplishes this at

time A and B are $\Psi_{A,S_1}(X_2)$ and $\Psi_{B,S_1}(X_2)$, respectively, where:

$$\Psi_{g,S_1}(X_2) = \frac{\Pr_g(X_1 = 0)}{\Pr_g(X_1 = 0|X_2)}, \quad g = A, B.$$

Using these reweighting terms, we can write the counterfactual distribution of wages that would have prevailed in the absence of unions as:

$$F_{Y_g^{C,S_1}}(y) = \int\int F_{Y_g|X_g}(y|X_1 = 0, X_2)\ \Psi_{g,S_1}(X_2)\, \mathrm{d}F_{X_g}(X_2, X_1 = 0), \quad g = A, B.$$

These various counterfactual distributions can then be used to compute the contribution of unions (or another binary variable x_1) to the composition effect, $\Delta_{X_1}^{F(y)}$, and to the wage structure effect, $\Delta_{S_1}^{F(y)}$:

$$\Delta_{X_1}^{F(y)} = F_{Y_A}(y) - F_{Y_A^{C,X_1}}(y), \tag{39}$$

and

$$\Delta_{S_1}^{F(y)} = [F_{Y_A}(y) - F_{Y_A^{C,S_1}}(y)] - [F_{Y_B}(y) - F_{Y_B^{C,S_1}}(y)]. \tag{40}$$

Although we need three different reweighting factors ($\Psi_{X_1}(X_1, X_2)$, $\Psi_{A,S_1}(X_2)$, and $\Psi_{B,S_1}(X_2)$) to compute the elements of the detailed wage decomposition corresponding to X_1, these three reweighting factors can be constructed from the estimates of the two probability models $\Pr_A(X_1 = 1|X_2)$ and $\Pr_B(X_1 = 1|X_2)$. As before, once these reweighting factors have been computed, the different counterfactual statistics are easily obtained using standard statistical packages.

General covariates

It is difficult to generalize the approach suggested above to the case of covariates that are not binary. In the case of the composition effect, one approach that has been followed in the applied literature consists of sequentially adding covariates in the probability model $\Pr(D_B = 1|X)$ used to compute $\Psi(X)$.[65] For instance, start with $\Pr(D_B = 1|X_1)$, compute $\Psi_1(X_1)$ and the counterfactual statistics of interest by reweighting. Then do the same thing with $\Pr(D_B = 1|X_1, X_2)$, etc.

One shortcoming of this approach is that the results depend on the order in which the covariates are sequentially introduced, just like results from a sequential decomposition

[65] See, for example, Butcher and DiNardo (2002) and Altonji et al. (2008).

for the mean also depend on the order in which the covariates are introduced in the regression. For instance, estimates of the effect of unions that fail to control for any other covariates may be overstated if union workers tend to be concentrated in industries that would pay high wages even in the absence of unions. As pointed out by Gelbach (2009), the problem with sequentially introducing covariates can be thought of as an omitted variable problem. Unless there are compelling economic reasons for first looking at the effect of some covariates without controlling for the other covariates, sequential decompositions will have the undesirable property of depending (strongly in some cases) on the order of the decomposition (path dependence).[66]

Fortunately, there is a way around the problem of path dependence when performing detailed decompositions using reweighting methods. The approach however still suffers from the adding-up problem and is more appropriate when only the effect of a particular factor is of interest. To illustrate this approach, consider a case with three covariates X_1, X_2, and X_3. In a sequential decomposition, one would first control for X_1 only, then for X_1 and X_2, and finally for X_1, X_2, and X_3. On the one hand, the regression coefficient on X_1 and/or X_2 in regressions that fail to control for X_3 are biased because of the omitted variable problem. The corresponding elements of a detailed OB decomposition for the mean based on these estimated coefficients would, therefore, be biased too.

On the other hand, the coefficient on the last covariate to be introduced in the regression (X_3) is not biased since the other covariates (X_1 and X_2) are also controlled for. So although order matters in a sequential regression approach, the effect of the last covariate to be introduced is not affected by the omitted variable bias.

The same logic applies in the case of detailed decompositions based on a reweighting approach. Intuitively, the difference in the counterfactual distribution one gets by reweighting with X_1 and X_2 only, comparing to reweighting with X_1, X_2, *and* X_3 should yield the appropriate contribution of X_3 to the composition effect.

To see this more formally, consider the group A counterfactual distribution that would prevail if the distribution of X_3, conditional on X_1, X_2, was as in group B:

$$
\begin{aligned}
F_{Y_A^{C,X_3}}(y) &= \int F_{Y_A|X_A}(y|X)\mathrm{d}F_{X_B}(X_3|X_1, X_2)\mathrm{d}F_{X_A}(X_1, X_2), \\
&= \int F_{Y_A|X_A}(y|X)\Psi_{X_3|X_1,X_2}(X_1, X_2)\mathrm{d}F_{X_A}(X_3|X_1, X_2)\mathrm{d}F_{X_A}(X_1, X_2), \\
&= \int F_{Y_A|X_A}(y|X)\Psi_{X_3|X_1,X_2}(X_1, X_2)\mathrm{d}F_{X_A}(X_1, X_2, X_3),
\end{aligned}
$$

[66] Both Butcher and DiNardo (2002) and Altonji et al. (2008) consider cases where there is indeed a good reason for following a particular order in the decomposition. For instance, Altonji et al. (2008) argue that, when looking at various youth outcomes, one should first control for predetermined factors like gender and race before controlling for other factors determined later in life (AFQT score, educational achievement, etc.). In such a situation, the decomposition is econometrically interpretable even if gender and race are introduced first without controlling for the other factors.

where the reweighting factor $\Psi_{X_3|X_1,X_2}(X_1, X_2)$ can be written as:

$$\begin{aligned}\Psi_{X_3|X_1,X_2}(X_1, X_2) &\equiv \frac{\mathrm{d}F_{X_B}(X_3|X_1, X_2)}{\mathrm{d}F_{X_A}(X_3|X_1, X_2)} \\ &= \frac{\mathrm{d}F_{X_B}(X_1, X_2, X_3)/\mathrm{d}F_{X_B}(X_1, X_2)}{\mathrm{d}F_{X_A}(X_1, X_2, X_3)/\mathrm{d}F_{X_A}(X_1, X_2)} \\ &= \Psi(X_1, X_2, X_3)/\Psi_{X_1,X_2}(X_1, X_2).\end{aligned}$$

$\Psi(X_1, X_2, X_3)$ is the reweighting factor used to compute the aggregate decomposition in Section 4.5. $\Psi_{X_1,X_2}(X_1, X_1)$ is a reweighting factor based on all the covariates except the one considered for the detailed decomposition (X_3). As before, Bayes' rule can be used to show that:

$$\Psi_{X_3|X_1,X_2}(X_1, X_2) = \frac{\Pr(X_1, X_2|D_B = 1)}{\Pr(X_1, X_2|D_B = 0)} = \frac{\Pr(D_B = 1|X_1, X_2)/\Pr(D_B = 1)}{\Pr(D_B = 0|X_1, X_2)/\Pr(D_B = 0)}.$$

Once again, this new reweighting factor is easily computed by running a probit or logit regression (with X_1 and X_2 as covariates) and using predicted probability to estimate $\Psi_{X_3|X_1,X_2}(X_1, X_2)$.

This reweighting procedure for the detailed decomposition is summarized as follows:

1. Compute the reweighting factor using all covariates, $\Psi(X)$.
2. For each individual covariate k, compute the reweighting factor using all covariates but X_k, $\Psi_{X_{-k}}(X_{-k})$.
3. For each covariate k, compute the counterfactual statistic of interest using the ratio of reweighting factors $\Psi(X)\Psi_{X_{-k}}(X_{-k})$ as weight, and compare it to the counterfactual statistic obtained using only $\Psi(X)$ as weight. The difference is the estimated contribution of covariate k to the composition effect.

Note that while this procedure does not suffer from path dependence, the contribution of each covariates does not sum up to the total contribution of covariates (aggregate composition effect). The difference is an interaction effect between the different covariates which is harder to interpret.

Advantages

This reweighting procedure shares most of the advantages of the other reweighting procedures we proposed for the aggregate decomposition. First, it is generally easy to implement in practice. Second, by using a flexible specification for the logit/probit, it is possible to get estimates of the various components of the decomposition that depend minimally on functional form assumptions. Third, the procedure yields efficient estimates.

Limitations

With a large number of covariates, one needs to compute a sizable number of reweighting factors to compute the various elements of the detailed decomposition. This can be tedious, although it does not require that much in terms of computations since each probit/logit is easy to estimate. Another disadvantage of the suggested decomposition is that although it does not suffer from the problem of path dependence, we are still left with an interaction term which is difficult to interpret. For these reasons, we suggest to first use a regression-based approach like the RIF-regression approach discussed above, which is essentially as easy to compute as a standard OB decomposition. The reweighting procedure suggested here can then be used to probe these results, and make sure they are robust to the functional-form assumptions implicit in the RIF-regression approach.

5.4. Detailed decomposition based on conditional quantiles

As we mentioned earlier, the method of Machado and Mata (2005) can be used to compute the wage structure sub-components of the detailed decomposition. These components are computed by sequentially switching the coefficients of the quantile regressions for each covariate from their estimated valued for group B to their estimated values for group A. This sequential switching cannot be used, however, to compute the sub-components of the composition effect of the detailed decomposition. Rather, Machado and Mata (2005) suggest an unconditional reweighting approach to do so. This does not provide a consistent effect since the effect of the reweighted covariate of interest gets confounded by other covariates correlated with that same covariate. For instance, if union workers are more concentrated in manufacturing, doing an unconditional reweighting on unions will also change the fraction of workers in manufacturing. In this sense the effect of unions is getting confounded by the effect of manufacturing.

This is a significant drawback since it is arguably more important to conduct a detailed decomposition for the composition effect than for the wage structure effect. As discussed earlier, there are always some interpretation problems with the detailed components of the wage structure effect because of the omitted group problem.

One solution is to use the conditional reweighting procedure described above instead. But once this type of reweighting approach is used, there is no need to estimate (conditional) quantile regressions. Unless the quantile regressions are of interest on their own, it is preferable to use a more consistent approach, such as the one based on the estimation of RIF-regressions, for estimating the detailed components of both the wage structure and composition effects.

6. EXTENSIONS

In this section, we present three extensions to the decomposition methods discussed earlier. We first consider the case where either the ignorability or the zero conditional mean assumptions are violated because of self-selection or endogeneity of the covariates.

We next discuss the situation where some of these problems can be addressed when panel data are available. We conclude the section by discussing the connection between conventional decomposition methods and structural modeling.

6.1. Dealing with self-selection and endogeneity

The various decomposition procedures discussed up to this point provide consistent estimates of the aggregate composition and wage structure effects under the ignorability assumption. Stronger assumptions, such as conditional mean independence (for decompositions of the mean) or straight independence, have to be invoked to perform the detailed decomposition. In this section we discuss some alternatives for estimating the decomposition when these assumptions fail. We mostly focus on the case of the OB decomposition of the mean, though some of the results we present could be extended to more general distributional statistics.

We consider three scenarios, first introduced in Section 2.1.6, under which the OB decomposition is inconsistent because of a failure of the ignorability or conditional independence assumption. In the first case, the problem is that individuals from groups A and B may self-select differently into the labor market. For instance, participation decisions of men (group B) may be different from participation decisions of women (group A) in ways that are not captured by observable characteristics. In the second case, we consider what happens when individuals can self-select into group A or B (for instance union and non-union jobs) on the basis of unobservables. The third case is a standard endogeneity problem where the covariates are correlated with the error term. For example, education (one of the covariate) may be correlated with the error term because more able individuals tend to get more schooling.

1. Differential self-selection within groups A and B.

One major concern when decomposing differences in wages between two groups with very different labor force participation rates is that the probability of participation depends on unobservables ε in different ways for groups A and B. This is a well known problem in the gender wage gap literature (Blau and Kahn, 2006; Olivetti and Petrongolo, 2008; Mulligan and Rubinstein, 2008, etc.) and in the black-white wage gap literature (Neal and Johnson, 1996).

Our estimates of decomposition terms may be directly affected when workers of groups A and B self-select into the labor market differently. Thus, controlling for selection based on observables and unobservables is necessary to guarantee point identification of the decomposition terms. If no convincing models for self-selection is available a more agnostic approach based on bounds has also been recently proposed. Therefore, following Machado (2009), we distinguish three branches in the literature of self-selection: (i) selection on observables; (ii) selection based on unobservables; (iii) bounds.

Selection based on observables and, when panel data are available, on time-invariant unobserved components can be used to impute values for the missing data on wages of non-participants. Representative papers of this approach are Neal and Johnson (1996), Johnson et al. (2000), Neal (2004), Blau and Kahn (2006) and Olivetti and Petrongolo (2008). These papers are typically concerned with mean or median wages. However, extensions to cumulative distribution functions or general ν-wage gaps could also be considered.

When labor market participation is based on unobservables, correction procedures for the mean wages are also available. In these procedures, a control variate is added as a regressor in the conditional expectation function. The exclusion restriction that an available instrument Z does not belong to the conditional expectation function also needs to be imposed.[67] Leading parametric and nonparametric examples are Heckman (1974, 1976), Duncan and Leigh (1980), Dolton and Makepeace (1986), Vella (1998), Mulligan and Rubinstein (2008).

In this setting, the decomposition can be performed by adding a control variate $\lambda_g(X_i, Z_i)$ to the regression. In most applications, $\lambda_g(X_i, Z_i)$ is the usual inverse Mills' ratio term obtained by fitting a probit model of the participation decision. Note that the addition of this control variate slightly changes the interpretation of the decomposition. The full decomposition for the mean is now

$$\begin{aligned}\Delta^{\mu} = (\beta_{B0} - \beta_{A0}) + \sum_{k=1}^{K} \overline{X}_{Bk} \left(\beta_{Bk} - \beta_{Ak}\right) + \overline{\lambda}_B(\sigma_B - \sigma_A) \\ + \sum_{k=1}^{K} \left(\overline{X}_{Bk} - \overline{X}_{Ak}\right) \beta_{Ak} + \left(\overline{\lambda}_B - \overline{\lambda}_A\right) \sigma_A.\end{aligned}$$

where σ_A and σ_B are the estimated coefficients on the control variates. The decomposition provides a full accounting for the wage gap that also includes differences in both the composition of unobservables $((\overline{\lambda}_B - \overline{\lambda}_A)\sigma_A)$ and in the return to unobservables $(\overline{\lambda}_B(\sigma_B - \sigma_A))$. This treats symmetrically the contribution of observables (the X's) and unobservables in the decomposition.

A third approach uses bounds for the conditional expectation function of wages for groups A and B. With those bounds one can come up with bounds for the wage structure effect, Δ_S^{μ}, and the composition effect, Δ_X^{μ}. Let $\Delta_S^{\mu} = \mathbb{E}[(\mathbb{E}[Y_B|X, D_B = 1] - \mathbb{E}[Y_A|X, D_B = 1])|D_B = 1]$. Then, letting D_S be a dummy indicating labor force participation, we can write the conditional expected wage as

$$\mathbb{E}\left[Y_g|X, D_g\right] = \mathbb{E}\left[Y_g|X, D_g, D_S = 0\right] + \Pr\left(D_S = 1|X, D_g\right)$$

[67] As is well known, selection models can be identified on the basis of functional restrictions even when an excluded instrumental variable is not available. This is no longer viewed, however, as a credible identification strategy. We, therefore, only focus on the case where an instrumental variable is available.

$$\times \left(\mathbb{E}\left[Y_g|X, D_g, D_S = 1\right] - \mathbb{E}\left[Y_g|X, D_g, D_S = 0\right]\right)$$

and therefore

$$\begin{aligned} & L_g + \Pr\left(D_S = 1|X, D_g\right)\left(\mathbb{E}\left[Y_g|X, D_g, D_S = 1\right] - L_g\right) \\ & \quad \leq \mathbb{E}\left[Y_g|X, D_g\right] \\ & \quad \leq U_g + \Pr\left(D_S = 1|X, D_g\right)\left(\mathbb{E}\left[Y_g|X, D_g, D_S = 1\right] - U_g\right) \end{aligned}$$

where L_g and U_g are lower and upper bounds of the distribution of Y_g, for $g = A, B$. Therefore,

$$\begin{aligned} & \left(\mathbb{E}\left[Y_B|X, D_B = 1, D_S = 1\right] - \mathbb{E}\left[Y_A|X, D_B = 1, D_S = 1\right]\right)\Pr\left(D_S = 1|X, D_B = 1\right) \\ & \quad + \left(L_B - U_A\right)\Pr\left(D_S = 0|X, D_B = 1\right) \\ & \leq \mathbb{E}\left[Y_B|X, D_B = 1\right] - \mathbb{E}\left[Y_A|X, D_B = 1\right] \\ & \leq \left(\mathbb{E}\left[Y_B|X, D_B = 1, D_S = 1\right]\right. \\ & \quad \left. - \mathbb{E}\left[Y_A|X, D_B = 1, D_S = 1\right]\right)\Pr\left(D_S = 1|X, D_B = 1\right) \\ & \quad + \left(U_B - L_A\right)\Pr\left(D_S = 0|X, D_B = 1\right). \end{aligned}$$

This bounding approach to the selection problem may also use restrictions motivated by econometric or economic theory to narrow the bounds, as in Manski (1990) and Blundell et al. (2007).

2. Self-Selection into groups A and B

In the next case we consider, individuals have the choice to belong to either group A or B. The leading example is the choice of the union status of workers. The traditional way of dealing with the problem is to model the choice decision and correct for selection biases using control function methods.[68]

As discussed in Section 2.1.6, it is also possible to apply instrumental variable methods more directly without explicitly modeling the selection process into groups A and B. Imbens and Angrist (1994) show that this will identify the wage gap for the subpopulation of compliers who are induced by the instrument to switch from one group to the other.

3. Endogeneity of the covariates

The standard assumption used in the OB decomposition is that the outcome variable Y is linearly related to the covariates, X, and that the error term υ is conditionally independent of X, as in Eq. (1). Now consider the case where the conditional independence assumption fails because one or several of the covariates are correlated with the

[68] See for instance, the survey of Lewis (1986) who concludes that these methods yield unreliable estimates of the union wage gap. Given these negative results and the lack of credible instruments for unionization, not much progress has been made in this literature over the last two decades. One exception is DiNardo and Lee (2004) who use a regression discontinuity design.

error term. Note that while the ignorability assumption may hold even if conditional independence fails, we consider a general case here where neither assumption holds.

As is well known, the conventional solution to the endogeneity problem is to use instrumental variable methods. For example, if we suspect years of education (one of the covariate) to be correlated with the error term in the wage equation, we can still estimate the model consistently provided that we have a valid instrument for years of education. The decomposition can then be performed by replacing the OLS estimates of the β coefficients by their IV counterparts.

Of course, in most cases it is difficult to come up with credible instrumentation strategies. It is important to remember, however, that even when the zero conditional mean assumption $\mathbb{E}(\upsilon|X) = 0$ fails, the aggregate decomposition may remain valid, provided that ignorability holds. This would be the case, for example, when unobserved ability is correlated with education, but the correlation (more generally the conditional distribution of ability given education) is the same in group A and B. While we are not able to identify the contribution of education vs. ability in this context (unless we have an instrument), we know that there are no systematic ability differences between groups A and B once we have controlled for education. As a result, the aggregate decomposition remains valid.

6.2. Panel data

An arguably better way of dealing with the selection and endogeneity problems mentioned above is to use panel data. Generally speaking, panel data methods can be used to compute consistent estimates of the β's in each of the three cases discussed earlier. For example, if the zero conditional mean assumption holds once we also control for a person-specific fixed effects θ_i in a panel of length T ($\mathbb{E}(\upsilon_{it}|X_{i1}, \ldots, X_{iT}, \theta_i) = 0$), we can consistently estimate β using standard panel data methods (fixed effects, first differences, etc.). This provides an alternative way of dealing with endogeneity problems when no instrumental variables are available.

As we also discussed earlier, panel data can be used to impute wages for years where an individual is not participating in the labor market (e.g. Olivetti and Petrongolo, 2008). Note that in cases where groups are mutually exclusive (e.g. men vs. women), it may still be possible to estimate fixed effect models if the basic unit used is the firm (or related concepts) instead (Woodcock, 2008). Care has to be exercised in those circumstances to ensure that the firm fixed effect is the same for both female and male employees of the same firm. Another important issue with these models is the difficulty of interpretation of the differences in male and female intercepts which may capture the unobserved or omitted individual and firm effects.

Panel data methods have also been used to adjust for the selection into groups in cases where the same individual is observed in group A and B. For example, Freeman (1984) and Card (1996) estimate the union wage gap with panel data to control for the selection

of workers into union status. Lemieux (1998) uses a more general approach where the return to the fixed effect may be different in the union and non-union sector. He also shows how to generalize the approach to the case of a decomposition of the variance.

Without loss of generality, assume that the return to the fixed effect for non-union workers (group A) is 1, while it is equal to σ_B for union workers. The mean decomposition adjusted for fixed effects yields:

$$\Delta^{\mu} = (\beta_{B0} - \beta_{A0}) + \sum_{k=1}^{K} \overline{X}_{Bk} (\beta_{Bk} - \beta_{Ak}) + \overline{\theta}_B(\sigma_B - 1) + \sum_{k=1}^{K} \left(\overline{X}_{Bk} - \overline{X}_{Ak}\right) \beta_{Ak} + \left(\overline{\theta}_B - \overline{\theta}_A\right).$$

The interpretation of the decomposition is the same as in a standard OB setting except that $(\overline{\theta}_B - \overline{\theta}_A)$ now represents the composition effect term linked to non-random selection into the union sector, while the wage structure term $\overline{\theta}_B(\sigma_B - 1)$ captures a corresponding wage structure effect.

More sophisticated models with several levels of fixed effects have also been used in practice. For instance, Abowd et al. (2008) decompose inter-industry wage differentials into various components that include both individual- and firm-specific fixed effects.

6.3. Decomposition in structural models

In Section 2, we pointed out that decomposition methods were closely related to methods used in the program evaluation literature where it is not necessary to estimate a fully specified structural model to estimate the main parameter of interest (the *ATT*). Provided that the ignorability assumption is satisfied, we can perform an aggregate decomposition without estimating an underlying structural model.

There are some limits, however, to what can be achieved without specifying any structure to the underlying economic problem. As we just discussed in Section 6.1, one problem is that the ignorability assumption may not hold. Under this scenario, more explicit modeling may be useful for correcting biases in the decomposition due to endogeneity, self-selection, etc.

Another problem that we now address concerns the interpretation of the wage structure components of the detailed decomposition. Throughout this chapter, we have proposed a number of ways of estimating these components for both the mean and more general distributional statistics. In the case of the mean, the interpretation of the detailed decomposition for the wage structure effect is relatively straightforward. Under the assumption (implicit in the OB decomposition) that the wage equations are truly linear and the errors have a zero conditional mean, we can think of the wage setting model as a fully specified structural model. The β coefficients are the "deep" structural

parameters of the model, and these structural parameters are used directly to perform the decomposition.

Things become more complicated once we go beyond the mean. For instance, in the case of the variance (Section 4.1), recall that the wage structure effect from Eq. (26) which depends on the parameters of both the models for the conditional mean (β) and for the variance (δ).

Take, for example, the case where one of the covariates is the union status of workers. The parameter δ captures the "compression", or within-group, effect, while the parameter β captures the "wage gap", or between-group, effect. These two terms have a distinct economic interpretation as they reflect different channels through which union wage policies tend to impact the wage distribution.

In the case of more general distributional statistics, the wage structure effect depends on an even larger number of underlying parameters capturing the relationship between the covariates and higher order moments of the distribution. As a result, the wage structure part of the detailed decomposition becomes even harder to interpret, as it potentially depends on a large number of underlying parameters.

In some cases, this may not pose a problem from an interpretation point of view. For instance, we may only care about the overall effect of unions, irrespective of whether it is coming from a between- or within-group effect (or corresponding components for higher order moments). But in other cases this type of interpretation may be unsatisfactory. Consider, for example, the effect of education on the wage structure. Like unions, education may influence wage dispersion through a between- or within-group channel. The between-group component is linked to the traditional return to education (effect on conditional means), but education also has a substantial effect on within-group dispersion (see, e.g., Lemieux, 2006b). All these effects are combined together in the decomposition methods proposed in Section 5, which is problematic if we want to know, for instance, the specific contribution of changes in the return to education to the growth in wage inequality.

In these circumstances, we need to use a more structural approach to get a more economically interpretable decomposition of the wage structure effect. The decomposition method of Juhn et al. (1993) is, in fact, an early example of a more structurally-based decomposition. In their setting, the model for the conditional mean is interpreted as an underlying human capital pricing equation. Likewise, changes in residual wage dispersion (given X) are interpreted as reflecting an increase in the return to unobservable skills.

As we discussed in Section 4.3, the fact that Juhn et al. (1993) provides a richer interpretation of the wage structure effect by separating the within- and between-group components is an important advantage of the method. We also mentioned, however, that the interpretation of the decomposition was not that clear for distributional statistics going beyond the variance, and that the procedure typically imposes substantial restrictions on the data that may or may not hold. By contrast, a method like DFL imposes very

little restrictions (provided that the probit/logit model used for reweighting is reasonably flexible), though it is more limited in terms of the economic interpretation of the wage structure effect.

In light of this, the challenge is to find a way of imposing a more explicit structure on the economic problem while making sure the underlying model "fits" the data reasonably well. One possible way of achieving this goal is to go back to the structural form introduced in Section 2 ($Y_{gi} = m_g(X_i, \varepsilon_i)$), and use recent results from the literature on nonparametric identification of structural functions to identify the functions $m_g(\cdot)$. As discussed in Section 2.2.1, this can be done by invoking results obtained by Matzkin (2003), Blundell and Powell (2007) and Imbens and Newey (2009). Generally speaking, it is possible to identify the functions $m_g(\cdot)$ nonparametrically under the assumptions of independence of ε (Assumption 8), and strict monotonicity of $m_g(\cdot)$ in ε (Assumption 9).

But while it is possible, in principle, to nonparametrically identify the functions $m_g(\cdot)$, there is no guarantee that the resulting estimates will be economically interpretable. As a result, a more common approach used in the empirical literature is to write down a more explicit (and parametric) structural model, but carefully look at whether the model adequately fits the data. Once the model has been estimated, simulation methods can then be used to compute a variety of counterfactual exercises. The counterfactuals then form the basis of a more economically interpretable decomposition of the wage structure effect.

To take a specific example, consider the Keane and Wolpin (1997) model of career progression of young men, where educational and occupational choices are explicitly modeled using a dynamic programming approach. After carefully looking at whether the estimated model is rich enough to adequately fit the distribution of wages, occupational choices, and educational achievement, Keane and Wolpin use the estimated model to decompose the distribution of lifetime utility (itself computed using the model). They conclude that 90 percent of the variance of lifetime utility is due to skill endowment heterogeneity (schooling at age 16 and unobserved type). By contrast, choices and other developments happening after age 16 have a relatively modest impact on the variance of lifetime utility.[69] The general idea here is to combine structural estimation and simulation methods to quantify the contribution of the different parameters of interest to some decompositions of interest. These issues are discussed in more detail in the chapter on structural methods by Keane et al. (2011).

One last point is that the interpretation problem linked to the wage structure effect does not apply to the detailed decomposition for the composition effect. In that case, each component is based on a clear counterfactual exercise that does not require an underlying structure to be interpretable. The aggregate decomposition is based on the

[69] Note, however, that Hoffman (2009) finds that skill endowments have a sizably smaller impact in a richer model that incorporates comparative advantage (across occupations), search frictions, and exogenous job displacement.

following counterfactual exercise: what would be the distribution of outcomes for group *A* if the distribution of the covariates for group *A* were the same as for group *B*? Similarly, the detailed decomposition is based on a conditional version of the counterfactual. For example, one may want to ask what would be the distribution of outcomes for group *A* if the distribution of unionization (or another covariate) for group *A* was the same as for group *B*, *conditional* on the distribution of the other covariates remaining the same.

These interpretation issues aside, it may still be useful to use a more structural approach when we are concerned about the validity of the decomposition because of self-selection, endogeneity, etc. For instance, in Keane and Wolpin (1997), the choice of schooling and occupation is endogenous. Using standard decomposition methods to look, for instance, at the contribution of the changing distribution of occupations to changes in the distribution wages would yield invalid results because occupational choice is endogenous. In such a context, structural modeling, like the IV and selection methods discussed in Section 6.1, can help recover the elements of the decomposition when standard methods fail because of endogeneity or self-selection. But the problem here is quite distinct from issues with the wage structure effect where standard decomposition methods are limited because of an interpretation problem, and where structural modeling provides a natural way of resolving this interpretation problem. By contrast, solutions to the problem of endogeneity or self-selection are only as a good as the instruments (or related assumptions) used to correct for these problems. As a result, the value added of the structural approach is much more limited in the case of the composition effect than in the case of the wage structure effect.

This last point is very clear in the emerging literature where structural modeling is used in conjunction with experimental data. For example, Card and Hyslop (2005) use experimental data from the Self Sufficiency Project (SSP) to look at why individuals offered with a generous work subsidy are less likely to receive social assistance (SA). By definition, there is no composition effect since the treatment and control groups are selected by random assignment. In that context, the average treatment effect precisely corresponds to the wage structure effect (or "SA" structure effect in this context) in a decomposition of the difference between the treatment and control group. It is still useful, however, to go beyond this aggregate decomposition to better understand the mechanisms behind the measured treatment effect. Card and Hyslop (2005) do so by estimating a dynamic search model.

This provides much more insight into the "black box" of the treatment effect than what a traditional decomposition exercise would yield. Remember that the detailed wage structure component in a OB type decomposition is based on the difference between the return to different characteristics in the two groups. In a pure experimental context like the SSP project, this simply reflects some heterogeneity in the treatment effect across different subgroups. Knowing about the importance of heterogeneity in the treatment effect is important from the point of view of the generalizability of the results.

But unlike a structural approach, it provides relatively little insight on the mechanisms underlying the treatment effect.

7. CONCLUSION

The development of new decomposition methods has been a fertile area of research over the last 10-15 years. Building on the seminal work of Oaxaca (1973) and Blinder (1973), a number of procedures that go beyond the mean have been suggested and used extensively in practice. In this chapter, we have reviewed these methods and suggested a number of "best practices" for researchers interested in these issues. We have also illustrated how these methods work in practice by discussing existing applications and working through a set of empirical examples throughout the chapter.

Another important and recent development in this literature has linked decomposition methods to the large and growing literature on program evaluation and treatment effects. This connection is useful for several reasons. First, it helps clarify some interpretation issues with decompositions. In particular, results from the treatment effects literature can be used to show, for example, that we can give a structural interpretation to an aggregate decomposition under the assumption of ignorability. Another benefit of this connection is that formal results about the statistical properties of treatment effects estimators can also be directly applied to decomposition methods. This helps guide the choice of decomposition methods that have good statistical properties, and conduct inference on these various components of the estimated decomposition.

But this connection with the treatment effects literature also comes at a cost. While no structural modeling is required to perform a decomposition or estimate a treatment effect, these approaches leave open the question of what are the economic mechanisms behind the various elements of the decomposition (or behind the treatment effect). Now that the connection between decomposition methods and the treatment effects literature has been well established, an important direction for future research will be to improve the connection between decomposition methods and structural modeling.

The literature on inequality provides some useful hints on how this connection can be useful and improved upon. In this literature, decomposition methods have helped uncover the most important factors behind the large secular increase in the distribution of wages. Those include the return to education, de-unionization, and the decline in the minimum wage, to mention a few examples. These findings have spurred a large number of more conceptual studies trying to provide formal economic explanations for these important phenomena. In principle, these explanations can then be more formally confronted to the data by writing down and estimating a structural model, and using simulation methods to quantify the role of these explanations.

This suggest a two-step research strategy where "off-the-shelf" decomposition methods, like those discussed in this chapter, can first be used to uncover the main forces

underlying an economic phenomenon of interest. More "structural" decomposition methods could then be used to better understand the economics behind the more standard decomposition results. We expect such a research strategy to be a fruitful area of research in the years to come.

REFERENCES

Abowd, John M., Kramarz, Francis, Lengerman, Paul, Roux, Sebastien, 2008. Persistent inter-industry wage differences: rent sharing and opportunity costs. Working paper.

Albrecht, James, Björklund, Anders, Vroman, Susan, 2003. Is there a glass ceiling in Sweden? Journal of Labor Economics 21, 145–178.

Altonji, Joseph G., Blank, Rebecca, 1999. Race and gender in the labor market. In: Ashenfelter, O., Card., D. (Eds.), Handbook of Labor Economics, vol. 3C. Elsevier Science, Amsterdam.

Altonji, Joseph G., Matzkin, Rosa L., 2005. Cross section and panel data estimators for nonseparable models with endogenous regressors. Econometrica 73, 1053–1102.

Altonji, Joseph G., Bharadwaj, P., Lange, Fabian, 2008, Changes in the characteristics of American youth: Implications for adult outcomes. Working paper, Yale University.

Athey, Susan, Imbens, Guido W., 2006. Identification and inference in nonlinear difference-in-differences models. Econometrica 74, 431–497.

Autor, David H., Levy, Frank, Murnane, Richard, 2003. The skill content of recent technological change: an empirical exploration. Quarterly Journal of Economics 118, 1279–1333.

Autor, David H., Katz, Lawrence B., Kearney, Melissa S., 2005. Rising Wage Inequality: The Role of Composition and Prices. NBER Working Paper No. 11628, September.

Barsky, R., Bound, John, Charles, K., Lupton, J., 2002. Accounting for the black-white wealth gap: a nonparametric approach. Journal of the American Statistical Association 97, 663–673.

Bauer, Thomas K., Göhlmann, Silja, Sinning, Mathias, 2007. Gender differences in smoking behavior. Health Economics 19, 895–909.

Bauer, Thomas K., Sinning, Mathias, 2008. An extension of the Blinder–Oaxaca decomposition to nonlinear models. Advances in Statistical Analysis 92, 197–206.

Bertrand, Marianne, Hallock, Kevin F., 2001. The gender gap in top corporate jobs. Industrial and Labor Relations Review 55, 3–21.

Biewen, Martin, 2001. Measuring the effects of socio-economic variables on the income distribution: an application to the income distribution: an application to the East German transition process. Review of Economics and Statistics 83, 185–190.

Bitler, Marianne P., Gelbach, Jonah B., Hoynes, Hilary W., 2006. What mean impacts miss: distributional effects of welfare reform experiments. American Economic Review 96, 988–1012.

Black, Dan, Haviland, Amelia, Sanders, Seth, Taylor, Lowell, 2008. Gender wage disparities among the highly educated. Journal of Human Resources 43, 630–659.

Blau, Francine D., Kahn, Lawrence M., 1992. The gender earnings gap: learning from international comparisons. American Economic Review 82, 533–538.

Blau, Francine D., Kahn, Lawrence M., 1997. Swimming upstream: trends in the gender wage differential in the 1980s. Journal of Labor Economics 15, 1–42.

Blau, Francine D., Kahn, Lawrence M., 2003. Understanding international differences in the gender pay gap. Journal of Labor Economics 21, 106–144.

Blau, Francine D., Kahn, Lawrence M., 2006. The US gender pay gap in the 1990s: slowing convergence. Industrial & Labor Relations Review 60 (1), 45–66.

Blinder, Alan, 1973. Wage discrimination: reduced form and structural estimates. Journal of Human Resources 8, 436–455.

Blundell, Richard, Powell, James L., 2007. Censored regression quantiles with endogenous regressors. Journal of Econometrics 141, 65–83.

Blundell, Richard, Gosling, Amanda, Ichimura, Hidehiko, Meghir, Costas, 2007. Changes in the distribution of male and female wages accounting for employment composition using bounds. Econometrica 75, 323–363.

Bourguignon, Francois, 1979. Decomposable income inequality measures. Econometrica 47, 901–920.
Bourguignon, F., Ferreira, Francisco H.G., 2005. Decomposing changes in the distribution of household incomes: methodological aspects. In: Bourguignon, F., Ferreira, F.H.G., Lustig, N. (Eds.), The Microeconomics of Income Distribution Dynamics in East Asia and Latin America. World Bank, pp. 17–46.
Bourguignon, F., Ferreira, Francisco H.G., Leite, Philippe G., 2008. Beyond Oaxaca–Blinder: Accounting for differences in household income distributions. Journal of Economic Inequality 6, 117–148.
Busso, Matias, DiNardo, John, McCrary, Justin, 2009. New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators. IZA Discussion Paper No. 3998.
Butcher, Kristin F., DiNardo, John, 2002. The Immigrant and native-born wage distributions: evidence from United States censuses. Industrial and Labor Relations Review 56, 97–121.
Cain, Glen, 1986. The economic analysis of labor market discrimination: a survey. In: Ashenfelter, O.C., Layard, R. (Eds.), Handbook of Labor Economics, vol. 1. North-Holland, pp. 709–730.
Card, David, 1992. The Effects of Unions on the Distribution of Wages: Redistribution or Relabelling? NBER Working Paper 4195. National Bureau of Economic Research, Cambridge, Mass.
Card, David, 1996. The effect of unions on the structure of wages: a longitudinal analysis. Econometrica 64, 957–979.
Card, David, Hyslop, Dean R., 2005. Estimating the effects of a time-limited earnings subsidy for welfare-leavers. Econometrica 73, 1723–1770.
Chay, Kenneth Y., Lee, David S., 2000. Changes in relative wages in the 1980s: returns to observed and unobserved skills and black-white wage differentials. Journal of Econometrics 99 (1), 1–38.
Chernozhukov, Victor, Fernandez-Val, Ivan, Melly, Blaise, 2009. Inference on Counterfactual Distributions. CeMMAP working paper CWP09/09.
Chernozhukov, Victor, Fernandez-Val, Ivan, Galichon, A., 2010. Quantile and probability curves without crossing. Econometrica 78, 1093–1126.
Chiquiar, Daniel, Hanson, Gordon H., 2005. International migration, self-selection, and the distribution of wages: Evidence from Mexico and the United States. Journal of Political Economy 113, 239–281.
Cotton, Jeremiah, 1998. On the decomposition of wage differentials. Review of Economics and Statistics 70, 236–243.
Cowell, Frank A., 1980. On the structure of additive inequality measures. Review of Economic Studies 47, 521–531.
Denison, E.F., 1962. The sources of economic growth in the United States and the alternatives before us. Supplementary Paper No. 13. Committee for Economic Development, New York.
DiNardo, John, Fortin, Nicole M., Lemieux, Thomas, 1996. Labor market institutions and the distribution of wages, 1973-1992: a semiparametric approach. Econometrica 64, 1001–1044.
DiNardo, John, Lee, David S., 2004. Economic impacts of new unionization on private sector employers: 1984-2001. The Quarterly Journal of Economics 119, 1383–1441.
DiNardo, John, Lemieux, Thomas, 1997. Diverging male inequality in the United States and Canada, 1981-1988: do institutions explain the difference. Industrial and Labor Relations Review 50, 629–651.
Dolton, Peter John, Makepeace, Gerald H., 1986. Sample selection and male-female earnings differentials in the graduate labour market. Oxford Economic Papers 38, 317–341.
Doiron, Denise J., Riddell, W. Craig, 1994. The impact of unionization on male-female earnings differences in Canada. Journal of Human Resources 29, 504–534.
Donald, Stephen G., Green, David A., Paarsch, Harry J., 2000. Differences in wage distributions between Canada and the United States: an application of a flexible estimator of distribution functions in the presence of covariates source. Review of Economic Studies 67, 609–633.
Duncan, Gregory M., Leigh, Duane E., 1980. Wage determination in the union and nonunion sectors: a sample selectivity approach. Industrial and Labor Relations Review 34, 24–34.
Egel, Daniel, Graham, Bryan, Pinto, Cristine, 2009. Efficicient estimation of data combination problems by the method of auxiliary-to-study tilting. mimeo.
Even, William E., Macpherson, David A., 1990. Plant size and the decline of unionism. Economics Letters 32, 393–398.
Fairlie, Robert W., 1999. The absence of the African-American owned business: an analysis of the dynamics of self–employment. Journal of Labor Economics 17, 80–108.

Fairlie, Robert W., 2005. An extension of the Blinder-Oaxaca decomposition technique to logit and probit models. Journal of Economic and Social Measurement 30, 305–316.
Fields, Judith, Wolff, Edward N., 1995. Interindustry wage differentials and the gender wage gap. Industrial and Labor Relations Review 49, 105–120.
Firpo, Sergio, 2007. Efficient semiparametric estimation of quantile treatment effects. Econometrica 75, 259–276.
Firpo, Sergio, 2010. Identification and Estimation of Distributional Impacts of Interventions Using Changes in Inequality Measures. EESP-FGV. mimeo.
Firpo, Sergio, Fortin, Nicole M., Thomas, Lemieux, 2007. Decomposing Wage Distributions using Recentered Influence Functions Regressions. mimeo, University of British Columbia.
Firpo, Sergio, Fortin, Nicole M., Lemieux, Thomas, 2009. Unconditional quantile regressions. Econometrica 77 (3), 953–973.
Fitzenberger, Bernd, Kohn, Karsten, Wang, Qingwei, 2010. The erosion of union membership in Germany: determinants, densities, decompositions. Journal of Population Economics (forthcoming).
Foresi, Silverio, Peracchi, Franco, 1995. The conditional distribution of excess returns: an empirical analysis. Journal of the American Statistical Association 90, 451–466.
Fortin, Nicole M., Lemieux, Thomas, 1998. Rank regressions, wage distributions, and the gender gap. Journal of Human Resources 33, 610–643.
Fortin, Nicole M., 2008. The gender wage gap among young adults in the United States: the importance of money vs. people. Journal of Human Resources 43, 886–920.
Freeman, Richard B., 1980. Unionism and the dispersion of wages. Industrial and Labor Relations Review 34, 3–23.
Freeman, Richard B., 1984. Longitudinal analysis of the effect of trade unions. Journal of Labor Economics 2, 1–26.
Freeman, Richard B., 1993. How much has deunionization contributed to the rise of male earnings inequality?. In: Danziger, Sheldon, Gottschalk, Peter (Eds.), Uneven Tides: Rising Income Inequality in America. Russell Sage Foundation, New York, 133-63.
Frolich, Markus, 2004. Finite-sample properties of propensity-score matching and weighting estimators. Review of Economics and Statistics 86, 77–90.
Gardeazabal, Javier, Ugidos, Arantza, 2004. More on the identification in detailed wage decompositions. Review of Economics and Statistics 86, 1034–57.
Gelbach, Jonah B., 2002. Identified Heterogeneity in Detailed Wage Decompositions. mimeo, University of Maryland at College Park.
Gelbach, Jonah B., 2009. When Do Covariates Matter? And Which Ones, and How Much? mimeo, Eller College of Management, University of Arizona.
Gomulka, Joanna, Stern, Nicholas, 1990. The employment of married women in the United Kingdom, 1970–1983. Economica 57, 171–199.
Gosling, Amanda, Machin, Stephen, Meghir, Costas, 2000. The changing distribution of male wages in the U.K,. Review of Economic Studies 67, 635–666.
Greene, William H., 2003. Econometric Analysis, 5th ed., Pearson Education, Upper Saddle River, NJ.
Heckman, James, 1974. Shadow prices, market wages and labor supply. Econometrica 42, 679–694.
Heckman, James, 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Annals of Economic and Social Measurement 5, 475–492.
Heckman, James, 1979. Sample selection bias as a specification error. Econometrica 47, 153–163.
Heckman, James J., Smith, Jeffrey, Clements, Nancy, 1997a. Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts. Review of Economic Studies 64 (4), 487–535.
Heckman, James J., Ichimura, Hidehiko, Todd, Petra, 1997b. Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. Review of Economic Studies 64, 605–654.
Heckman, James J., Ichimura, Hidehiko, Smith, Jeffrey, Todd, Petra, 1998. Characterizing selection bias using experimental data. Econometrica 66, 1017–1098.
Heywood, John S., Parent, Daniel, 2009. Performance Pay and the White-Black Wage Gap. mimeo, McGill University.

Hirano, Kiesuke, Imbens, Guido W., Ridder, Geert, 2003. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71, 1161–1189.
Holland, Paul W., 1986. Statistics and causal inference. Journal of the American Statistical Association 81 (396), 945–960.
Hoffman, Florian, 2009. An Empirical Model of Life-Cycle Earnings and Mobility Dynamics. University of Toronto, Department of Economics. mimeo.
Horrace, William, Oaxaca, Ronald L., 2001. Inter-industry wage differentials and the gender wage gap: an identification problem. Industrial and Labor Relations Review 54, 611–618.
Imbens, Guido W., Angrist, Joshua, 1994. Identification and estimation of local average treatment effects. Econometrica 62, 467–476.
Imbens, Guido W., Newey, Whitney K., 2009. Identification and estimation of triangular simultaneous equations models without additivity. Econometrica 77 (5), 1481–1512.
Jann, Ben, 2005. Standard errors for the Blinder-Oaxaca decomposition. German Stata Users' Group Meetings 2005. Available from http://repec.org/dsug2005/oaxaca_se_handout.pdf.
Jann, Ben, 2008. The Oaxaca-Blinder decomposition for linear regression models. Stata Journal 8, 435–479.
Jones, Frank Lancaster, 1983. On decomposing the wage gap: a critical comment on blinder's method. Journal of Human Resources 18, 126–130.
Johnson, William, Kitamura, Yuichi, Neal, Derek, 2000. Evaluating a simple method for estimating black-white gaps in median wages. American Economic Review 90, 339–343.
Jorgenson, D.W., Griliches, Z., 1967. The explanation of productivity change. Review of Economic Studies 34, 249–283.
Juhn, Chinhui, Murphy, Kevin M., Pierce, Brooks, 1991. Accounting for the slowdown in black-white wage convergence. In: Kosters, M.H. (Ed.), Workers and Their Wages: Changing Patterns in the United States. American Enterprise Institute, Washington.
Juhn, Chinhui, Murphy, Kevin M., Pierce, Brooks, 1993. Wage inequality and the rise in returns to skill. Journal of Political Economy 101, 410–442.
Keane, Michael P., Wolpin, Kenneth I., 1997. The career decisions of young men. Journal of Political Economy 105, 473–522.
Keane, Michael P., Todd, Petra E., Wolpin, Kenneth I., 2011. The structural estimation of behavioral models: discrete choice dynamic programming methods and applications. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 4A. Elsevier Science, Amsterdam, pp. 331–461.
Kendrick, John W., 1961. Productivity Trends in the United States. Princeton University Press, Princeton.
Kennedy, Peter, 1986. Interpreting dummy variables. Review of Economics and Statistics 68, 174–175.
Kline, Pat, 2009. Blinder-Oaxaca as a Reweighting Estimator. UC Berkeley mimeo.
Koenker, Roger, Bassett, G., 1978. Regression quantiles. Econometrica 46, 33–50.
Krueger, Alan B., Summers, Lawrence H., 1988. Efficiency wages and the inter-industry wage structure. Econometrica 56 (2), 259–293.
Krieg, John M., Storer, Paul, 2006. How much do students matter? applying the Oaxaca decomposition to explain determinants of adequate yearly progress. Contemporary Economic Policy 24, 563–581.
Lemieux, Thomas, 1998. Estimating the effects of unions on wage inequality in a panel data model with comparative advantage and non-random selection. Journal of Labor Economics 16, 261–291.
Lemieux, Thomas, 2002. Decomposing changes in wage distributions: a unified approach. The Canadian Journal of Economics 35, 646–688.
Lemieux, Thomas, 2006a. Post-secondary education and increasing wage inequality. American Economic Review 96, 195–199.
Lemieux, Thomas, 2006b. Increasing residual wage inequality: composition effects, noisy data, or rising demand for skill?. American Economic Review 96, 461–498.
Lewis, H.Gregg, 1963. Unionism and Relative Wages in the United States. University of Chicago Press, Chicago.
Lewis, H.Gregg, 1986. Union Relative Wage Effects: A Survey. University of Chicago Press, Chicago.
Neumark, David, 1988. Employers' discriminatory behavior and the estimation of wage discrimination. Journal of Human Resources 23, 279–295.
Machado, José F., Mata, José, 2005. Counterfactual decomposition of changes in wage distributions using quantile regression. Journal of Applied Econometrics 20, 445–465.

Machado, Cecilia, 2009. Selection, Heterogeneity and the Gender Wage Gap. Columbia University, Economics Department. mimeo.
Manski, Charles F., 1990. Nonparametric bounds on treatment effects. American Economic Review 80 (2), 319–323.
Matzkin, Rosa L., 2003. Nonparametric estimation of nonadditive random functions. Econometrica 71 (5), 1339–1375.
McEwan, P.J., Marshall, J.H., 2004. Why does academic achievement vary across countries? Evidence from Cuba and Mexico. Education Economics 12, 205–217.
Melly, Blaise, 2006. Estimation of counterfactual distributions using quantile regression. University of St. Gallen, Discussion Paper.
Melly, Blaise, 2005. Decomposition of differences in distribution using quantile regression. Labour Economics 12, 577–590.
Mulligan, Casey B., Rubinstein, Yona, 2008. Selection, investment, and women's relative wages over time. Quarterly Journal of Economics 123, 1061–1110.
Neal, Derek A., Johnson, W., 1996. The role of premarket factors in black-white wage differences. Journal of Political Economy 104, 869–895.
Neal, Derek A., 2004. The measured black-white wage gap among women is too small. Journal of Political Economy 112, S1-S28.
Ñopo, Hugo, 2008. Matching as a tool to decompose wage gaps. Review of Economics and Statistics 90, 290–299.
Oaxaca, Ronald, 1973. Male-female wage differentials in urban labor markets. International Economic Review 14, 693–709.
Oaxaca, Ronald L., Ransom, Michael R., 1994. On discrimination and the decomposition of wage differentials. Journal of Econometrics 61, 5–21.
Oaxaca, Ronald L., Ransom, Michael R., 1998. Calculation of approximate variances for wage decomposition differentials. Journal of Economic and Social Measurement 24, 55–61.
Oaxaca, Ronald L., Ransom, Michael R., 1999. Identification in detailed wage decompositions. Review of Economics and Statistics 81, 154–157.
Oaxaca, Ronald L., 2007. The challenge of measuring labor market discrimination against women. Swedish Economic Policy Review 14, 199–231.
Olivetti, Claudia, Petrongolo, Barbara, 2008. Unequal pay or unequal employment? a cross-country analysis of gender gaps. Journal of Labor Economics 26, 621–654.
O'Neill, June, O'Neill, Dave, 2006. What do wage differentials tell us about labor market discrimination?. In: Polachek, Soloman, Chiswich, Carmel, Rapoport, Hillel (Eds.), The Economics of Immigration and Social Policy. Research in Labor Economics 24, 293–357.
Reimers, Cornelia W., 1983. Labor market discrimination against hispanic and black men. Review of Economics and Statistics 65, 570–579.
Robins, James, Rotnizky, Andrea, Ping Zhao, Lue, 1994. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association 89, 846–866.
Rosenbaum, Paul R., Rubin, Donald B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.
Rosenbaum, Paul R., Rubin, Donald B., 1984. Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association 79, 516–524.
Rothe, Christoph, 2009. Nonparametric estimation of distributional policy effects. Journal of Econometrics 155, 56–70.
Shorrocks, Anthony F., 1980. The class of additively decomposable inequality measures. Econometrica 48, 613–625.
Shorrocks, Anthony F., 1984. Inequality decomposition by population subgroups. Econometrica 52, 1369–1385.
Shorrocks, Anthony F. 1999. Decomposition Procedures for Distributional Analysis: A Unified Framework Based on the Shapley Value. University of Essex, Department of Economics. mimeo.
Solow, Robert, 1957. Technical change and the aggreagate production function. Review of Economics and Statistics 39, 312–320.
Sohn, Ritae, 2008. The Gender Math Gap: Is It Growing? mimeo, SUNY Albany.

Vella, Frank, 1998. Estimating models with sample selection bias: a survey. Journal of Human Resources 33, 127–169.
Woodcock, Simon D., 2008. Wage differentials in the presence of unobserved worker, firm, and match heterogeneity?. Labour Economics 15, 772–794.
Yun, Myeong-Su, 2005. A simple solution to the identification problem in detailed wage decomposition. Economic Inquiry 43, 766–772. with Erratum, Economic Inquiry (2006), 44: 198.
Yun, Myeong-Su, 2008. Identification problem and detailed oaxaca decomposition: a general solution and inference. Journal of Economic and Social Measurement 33, 27–38.

CHAPTER 2

Field Experiments in Labor Economics☆

John A. List[*,1], **Imran Rasul**[**,2]

[*] Department of Economics, University of Chicago, 1126 East 59th Street, Chicago, IL 60637, USA

[**] Department of Economics, University College London, Drayton House, 30 Gordon Street, London WC1E 6BT, United Kingdom

Contents

☆ We gratefully acknowledge financial support from ELSE . We thank the editors, Orley Ashenfelter and David Card for comments. We thank Alec Brandon, David Herberich, Dana Krueger, Richard Murphy, Yana Peysakhovich and László Sándor for excellent research assistance. All errors remain our own.

1 Tel.: +1-773-702-9811; fax: +1-773-702-8490.

2 Tel: +44-207-679-5853; fax: +44-207-916-2775.

E-mail addresses: jlist@uchicago.edu (John A. List), i.rasul@ucl.ac.uk (Imran Rasul).

Handbook of Labor Economics, Volume 4a ISSN 0169-7218, DOI 10.1016/S0169-7218(11)00408-4

Abstract

We overview the use of field experiments in labor economics. We showcase studies that highlight the central advantages of this methodology, which include: (i) using economic theory to design the null and alternative hypotheses; (ii) engineering exogenous variation in real world economic environments to establish causal relations and learn about the underlying mechanisms; and (iii) engaging in primary data collection and often working closely with practitioners. To highlight the potential for field experiments to inform issues in labor economics, we organize our discussion around the individual life cycle. We therefore consider field experiments related to the accumulation of human capital, the demand and supply of labor, and behavior within firms, and close with a brief discussion of the nascent literature of field experiments related to household decision making.

JEL classification: C93; J01

Keywords: Field experiments; Labor economics

1. INTRODUCTION

This chapter overviews the burgeoning literature in field experiments in labor economics. The essence of this research method involves researchers engineering carefully crafted exogenous variation into real world economic environments, with the ultimate aim of identifying the causal relationships and mechanisms underlying them. This chapter describes this approach and documents how field experiments have begun to yield new insights for research questions that have long been studied by labor economists.

Given our focus on such long-standing questions, in no way do we attempt to do justice to the enormous literature to which field experiments are beginning to add. Our aim is rather to showcase specific field experiments that highlight what we view to be the central advantages of this methodology: (i) using economic theory to design the null and alternative hypotheses; (ii) engineering exogenous variation in real world economic

environments to establish causal relations and learn the mechanisms behind them; and (iii) engaging in primary data collection and often working closely with practitioners.

As with any research methodology in economics, of course, not every question will be amenable to field experiments. Throughout our discussion, we will bring to the fore areas in labor economics that remain relatively untouched by field experiments. In each case, we try to distinguish whether this is simply because researchers have not had opportunities to design field experiments related to such areas, or whether the nature of the research question implies that field experiments are not the best approach to tackle the problem at hand. Finally, even among those questions where field experiments can and have provided recent insights, we will argue that such insights are most enhanced when we are able to combine them with, or take inspiration from, other approaches to empirical work in economics. For example, a number of studies we describe take insights from laboratory environments to show the importance of non-standard preferences or behaviors in real world settings. A second class of papers combine field experimentation with structural estimation to measure potential behavioral responses in alternative economic environments, and ultimately to help design optimal policies to achieve any given objective.

The bulk of the chapter is dedicated to documenting the use and insights from field experiments in labor economics. To emphasize the relevance of field experiments to labor economics, we organize this discussion by following an individual over their life cycle. More specifically, we begin by considering field experiments related to early childhood interventions and the accumulation of human capital in the form of education and skills. We then consider research questions related to the demand and supply of labor, and labor market discrimination. We then move on to consider research questions related to behavior within firms: how individuals are incentivized within firms, and other aspects of the employment relationship. Finally, we conclude with a brief discussion of the nascent literature on field experiments related to household decision making.

We have chosen these topics selectively on the basis of where carefully designed field experiments have already been conducted. Within each field, we have decided to discuss a small number of papers using field experiments that, in our opinion, showcase the best of what field experiments can achieve. In no way is our discussion meant to be an exhaustive survey of the literature on field experiments in labor economics. The literature arising just in the past decade has grown too voluminous for even a tome to do it justice.

In each stage of the life cycle considered, wherever appropriate, we try to discuss: (i) the link between the design of field experiments and economic theory; (ii) the benefits of primary data collection that is inherent in field experimentation to further probe theory and distinguish between alternative hypotheses; and (iii) how reduced form effects identified from a field experiment can be combined with theory and structural modelling to make sample predictions and inform policy design.

In the remainder of this section, we place our later discussion into historical context by describing the experimental approach in science, and arguing how among economists, labor economists have for decades been at the forefront of exploiting and advancing experimental approaches to identify causal relationships. We then lay the groundwork for the discussion in later sections. We define the common elements at the heart of all field experiments, and then present a more detailed typology to highlight the subtle distinctions between various sub-types of field experiments. This approach allows us to discuss the advantages and disadvantages of field experiments over other forms of experimentation, such as large scale social experiments and laboratory based experiments.[3]

Our final piece of groundwork is to identify key trends in published research in labor economics over the past decade. This allows us to organize our later discussion more clearly in two dimensions. First, we think of nearly all research questions in labor economics as mapping to particular stages of an individual's life cycle. We therefore roughly organize research questions in labor economics into those relating to the accumulation of human capital, labor market entry and labor supply choices, behavior within firms, and household decision making. Second, we are able to focus in on those sub-fields in labor economics where extant field experiments have already begun to make inroads and provide new insights. In turn, this helps us make precise the types of research question field experiments are most amenable to, areas in which field experiments have been relatively under supplied, and those research questions that are better suited to alternative empirical methods.

1.1. The experimental approach in science

The experimental approach in scientific inquiry is commonly traced to Galileo Galilei, who pioneered the use of quantitative experiments to test his theories of falling bodies. Extrapolating his experimental results to the heavenly bodies, he pronounced that the services of angels were not necessary to keep the planets moving, enraging the Church and disciples of Aristotle alike. For his efforts, Galileo is now viewed as the Father of Modern Science. Since the Renaissance, fundamental advances making use of the experimental method in the physical and biological sciences have been fast and furious.[4]

Taking the baton from Galileo, in 1672 Sir Isaac Newton used experimentation to show that white light is equal to purity, again challenging the preachings of Aristotle. The experimental method has produced a steady stream of insights. Watson and Crick used data from Rosalind Franklin's X-ray diffraction experiment to construct a theory of the chemical structure of DNA; Rutherford's experiments shooting charged particles at a piece of gold foil led him to theorize that atoms have massive, positively charged

[3] In this Handbook, Charness and Kuhn (2011, 2010) provide a useful discussion of extant laboratory studies in the area of labor economics.

[4] For a more complete discussion see List and Reiley (2008).

nuclei; Pasteur rejected the theory of spontaneous generation with an experiment that showed that micro-organisms grow in boiled nutrient broth when exposed to the air, but not when exposed to carefully filtered air. Even though the experimental method produced a steady flow of important facts for roughly 400 years, the proper construction of a counterfactual control group was not given foundations until the early twentieth century.

1.1.1. An experimental cornerstone

In 1919, Ronald Fisher was hired at Rothamsted Manor to bring modern statistical methods to the vast experimental data collected by Lawes and Gilbert (Levitt and List, 2009). The data collection methods at Rothamsted Manor were implemented in the standard way to provide practical underpinnings for the ultimate purpose of agricultural research: to provide management guidelines. For example, one of the oldest questions in the area of agricultural economics relates to agricultural yields: what is the optimal application rate of fertilizer, seed, and herbicides?

In an attempt to modernize the experimental approach at Rothamsted, Fisher introduced the concept of randomization and highlighted the experimental tripod: the concepts of replication, blocking, and randomization were the foundation on which the analysis of the experiment was based (Street, 1990). Of course, randomization was the linchpin, as the validity of tests of significance stems from randomization theory.

Fisher understood that the goal of any evaluation method is to construct the proper counterfactual. Without loss of generality, define y_{i1} as the outcome for observational unit i with treatment, y_{i0} as the outcome for unit i without treatment. The treatment effect for plot i can then be measured as $\tau_i = y_{i1} - y_{i0}$. The major problem, however, is one of a missing counterfactual—plot i is not observed in both states. Fisher understood that methods to create the missing counterfactual to achieve identification of the treatment effect were invaluable, and his idea was to use randomization.

As Levitt and List (2009) discuss, Fisher's fundamental contributions were showcased in agricultural field experiments, culminating with the landmark 1935 book, *The Design of Experiments*, which was a catalyst for the actual use of randomization in controlled experiments. At the same time, Jerzy Neyman's work on agricultural experimentation showcased the critical relationship between experiments and survey design and the pivotal role that randomization plays in both (Splawa-Neyman, 1923a,b). Neyman's work continued in the area of sampling and culminated in his seminal paper, published in 1934. As Rubin (1990) notes, it is clear that randomization was "in the air" in the early 1920s, and the major influences of the day were by scholars doing empirical research on agricultural related issues. Clearly, such work revolutionized the experimental approach and weighs on experimental designs in all sciences today.

As emphasized throughout, we view field experimenters as being engaged in data generation, primary data collection, and data evaluation. Labor economists in particular

have been at the forefront of the use of experimental designs, as highlighted by the following two historic examples.

1.1.2. Early labor market field experiments at the Hawthorne plant

In the 1920s the Western Electric Company was the monopoly supplier of telephone equipment to AT&T. Western opted to have its main factory, the Hawthorne plant located in the suburbs of Chicago, be the main supplier for this important contract. The Hawthorne plant was considered to be one of the most advanced manufacturing facilities in America at the time, and employed roughly 35,000 people, mainly first- and second-generation immigrants (Gale, 2004). Always open to new techniques to improve efficiency and profitability, officials of Western were intrigued when the National Academy of Sciences expressed interest in a hypothesis put forth by electrical suppliers, who claimed that better lighting improved productivity.

The experimental exercises that resulted have few parallels within social science. The indelible footprint of these experiments laid the groundwork for a proper understanding of social dynamics of groups and employee relations in the workplace. Indeed, the data drawn from this research became the thrust of the human relations movement of the twentieth century, and represent the underpinnings of contemporary efforts of industry to motivate and deal with workers. In academia, the Hawthorne data spawned the development of a new field of study—Industrial Psychology—and remains an important influence on the manner in which scientists conduct experimental research today. Many of the issues raised in these studies are considered part of mainstream personnel economics, as discussed in the chapter in this Handbook on Human Resource Management practices by Bloom and Van Reenen (2011). In a later section of this chapter we review how a new generation of field experiments have provided new insights into these age old questions of behavior within firms.[5]

The first experiments executed at the Hawthorne plant have been famously denoted the "illumination experiments" because they varied the amount of light in the workplace. More specifically, between 1924 and 1927 the level of lighting was systematically changed for experimental groups in different departments (Mayo, 1933, pp. 55-56, Roethlisberger and Dickson, 1939, pp. 14-18, provide a more complete account). Workers in these departments were women who assembled relays and wound coils of wire, and their output was measured as units completed per unit of time.[6]

Discussions of these data have been widespread and have been an important influence on building the urban legend. For instance, Franke and Kaul (1978, p. 624) note that "Inexplicably worker output... generally increased regardless of increase or decrease in

[5] Frederick Taylor's seminal book, *The Principles of Scientific Management*, published in 1911, which creatively considered techniques to shorten task time, was also an important stimulus for the Industrial Psychology field.

[6] A relay was a switching device activated in the telephone exchange as each number was dialed, and was a fairly mind-numbing task: assemble a coil, armature, contact springs, and insulators by fastening them to a fixture with four screws. On average, it was roughly one minute's worth of work.

illumination." Yet, the only account of these experiments published at the time is Snow (1927), published in an engineering newsletter, and he argues that "The corresponding production efficiencies by no means followed the magnitude or trend of the lighting intensities. The output bobbed up and down without direct relation to the amount of illumination." Unfortunately, the article does not present data or any statistical analysis. Ever since, the literature has remained at a state of question since people thought that the data were lost. Indeed, an authoritative voice on this issue, Rice (1982) notes that "the original research data somehow disappeared." Gale (2004, p. 439) expresses similar thoughts concerning the illumination experiments: "these particular experiments were never written up, the original study reports were lost, and the only contemporary account of them derives from a few paragraphs in a trade journal" (Roethlisberger and Dickson, 1939; Gillespie, 1991).[7]

Using data preserved in two library archives Levitt and List (2010) dug up the original data from the illumination experiment, long thought to be destroyed. Their analysis of the newly found data reveals little evidence to support the existence of a Hawthorne effect as commonly described. Namely, there is no systematic evidence that productivity jumped whenever changes in lighting occurred. Alternatively, they do uncover some weak evidence consistent with more subtle manifestations of Hawthorne effects in the data. In particular, output tends to be higher when experimental manipulations are ongoing relative to when there is no experimentation. Also consistent with a Hawthorne effect is that productivity is more responsive to experimenter manipulations of light than naturally-occurring fluctuations.

As mysterious and legendary as the illumination experiments have become, it is fair to say that the second set of experiments conducted at the plant—the relay assembly experiments—have kept academics busy for years. Using an experimental area constructed for the illumination experiments, beginning in April 1927, researchers began an experiment meant to examine the effect of workplace changes upon productivity. In this case, the task was relay assembly.

The researchers began by secretly observing the women in their natural environment for two weeks, and then used various treatments, including manipulating the environment in such a way to increase and decrease rest periods, over different temporal intervals. While their design certainly did not allow easy assessment of clean treatment effects, the experimenters were puzzled by the observed pattern: output seemingly rose regardless of the change implemented. When output remained high after the researchers returned conditions to the baseline—output had risen from 2400 relays per week to nearly 3000 relays per week—management became interested in identifying the underlying mechanisms at work.

[7] There are many other evaluations as well. For example, a controversial article written by Bramel and Friend (1981), heavily laced with Marxist ideology, takes a conspiratorial view of industrial psychologists and argues that the Hawthorne effect is simply the result of "capitalist bias among modern industrial psychologists."

Western Electric subsequently brought in academic consultants, including Elton Mayo, in 1928. With Mayo's assistance, the experiments continued and by February of 1929, when productivity was at a startling rate of a new relay dropped down the chute every 40-50 seconds, the company besieged the five women with attention, besides "a new test room supervisor, an office boy, and a lady who helped with the statistics" others could be added: "an intermittent stream of other visitors or consultants: industrialists, industrial relations experts, industrial psychologists, and university professors." (Gale, 2004, p. 443). The experiment lasted until June 1932, when the women in the test room received their notices (except the exceptional worker, Jennie Sirchio, who worked in the office for a few months before being let go) after the stock market crash of October 24, 1932. The crash induced one in ten US phones to be disconnected in 1932, leading to a decrease in Western Electric's monopoly rents of more than 80%.

The five year experiment provided a wealth of data, and much of the Hawthorne Effect's statistical underpinnings are a direct result of the relay assembly experiment. Mayo's (1933) results concluded that individuals would be more productive when they knew they were being studied.[8] For this insight, Mayo came to be known as the "father of the Hawthorne effect", and his work lead to the understanding that the workplace was, importantly, a system that was first and foremost social, and composed of several interdependent parts. When we present a detailed typology of field experiments later in this section, we make precise a distinction between those field experiments in which agents are aware of their participation in an experiment, and those in which they are unaware of the exogenous manipulation of their economic environment.

Mayo stressed that workers are not merely at work to earn an honest wage for an honest day's effort, rather they are more prominently influenced by social demands, their need for attention, input to decision making, and by the psychological factors of the environment. The notion that workers effort and behavior are driven by more than the monetary rewards of work, is an idea that has received close scrutiny among the most recent generation of field experiments in firms, as reviewed later.

Clearly, Mayo argued, being the object of attention with the study induced a sense of satisfaction among workers that made them feel proud and part of a cohesive unit, generating greater productivity levels than could ever be imagined. Mayo's disciples, Frist Roethlisberger and William Dickson, another engineer at Western Electric, produced a detailed assessment that focused mainly on the relay experimental data (Roethlisberger and Dickson, 1939) and generated similar conclusions. Industrial psychology would soon

[8] Derivative of this path-breaking experiment were two experiments run alongside the relay experiment. Both were started in August of 1928; one was a second relay experiment, the other a mica splitting experiment. In the second relay experiment, five women workers were subjected to variations in a small group incentive program from August 1928 to March 1929. In the mica splitting experiment, the researchers began by secretly monitoring the output of five women at their regular department workstations. Their job was to split, measure, and trim mica chips that were to be used for insulation. After observing the workers secretly, they moved the women to a special test room where, unlike their cohorts, they received 10-minute rest breaks at 9:30 a.m. and 2:30 p.m.

find an important place in undergraduate and graduate curricula. Again in later sections, we provide examples of where field experiments have taken insights from psychology and laboratory environments to check for the existence and quantitative importance of such behaviors that are not encompassed within neoclassical economic models.

It is difficult to understate the importance of these findings, as they have served as the paradigmatic foundation of the social science of work (Franke and Kaul, 1978), providing a basis for an understanding of the economics of the workplace, and dramatically influenced studies in organizational development and behavior, leadership, human relations, and workplace design. The results also provide an important foundation for experimental work within the social sciences, including economics, where one must constantly be aware of the effects argued to be important in the Hawthorne relay experiment.[9]

1.1.3. Large-scale social experiments

A second period of interest directly related to field experiments in labor economics is the latter half of the twentieth century, during which government agencies conducted a series of large-scale social experiments.[10] In the US, social experiments can be traced to Heather Ross, an MIT economics doctoral candidate working at the Brookings Institution. As Levitt and List (2009) discuss, Ross wrote a piece titled "A Proposal for Demonstration of New Techniques in Income Maintenance", in which she suggested a randomly assigned social experiment to lend insights into the policy debate.

The experiment that resulted began in 1968 in five urban communities in New Jersey and Pennsylvania: Trenton, Paterson, Passaic, and Jersey City in NJ, and Scranton, PA and eventually became Ross' dissertation research ("An Experimental Study of the Negative Income Tax", which cost more than $5 million—exceeding $30 million in today's dollars). The idea behind the experiment was to explore the behavioral effects of negative income taxation, a concept first introduced by Milton Friedman and Robert Lampman, who was at the University of Wisconsin's poverty institute.[11] The experiment, which targeted roughly 1300 male-headed households who had at least one employable

[9] The success of the relay assembly experiments led to in-depth surveys (from 1928-1931) and one final experiment in the Hawthorne plant—the "bank wiring" experiment, designed by Mayo and others from 1931-1932. The researchers began by examining the productivity of 14 men who assembled telephone terminals. They then moved these men to a special test room, without introducing any other changes in work or pay conditions. Despite the move to a separate experimental setting, the men's output did not increase.

[10] This, and the subsequent subsections, draw from Harrison and List (2004), List (2006), and Levitt and List (2009). There are many definitions of social experiments in the economics literature. Ferber and Hirsch (1982, p. 7) define a social experiment in economics as ".... a publicly funded study that incorporates a rigorous statistical design and whose experimental aspects are applied over a period of time to one or more segments of a human population, with the aim of evaluating the aggregate economic and social effects of the experimental treatments." Greenberg and Shroder (2004) define a social experiment as having at least the following four features: (i) random assignment, (ii) policy intervention, (iii) follow-up data collection, and (iv) evaluation.

[11] As the Editors pointed out, the basic idea of a negative income tax was a part of the liberal party platform in the 1940s, and it is usually argued that it was designed by Juliet Rhys-Williams, an amazing advocate for women in that period.

person, experimentally varied both the guaranteed level of income and the negative tax rate (Ross, 1970). The guaranteed level of income ranged from 50% to 125% of the estimated poverty line income level for a family of four ($1650-$4125 in 1968 dollars) while the negative income tax rate ranged from 30% to 70%.[12] The experiment lasted three years. Families in both the control and treatment groups were asked to respond to questionnaires every three months during this time span, with the questions exploring issues such as family labor supply, consumption and expenditure patterns, general mobility, dependence on government, and social integration.

The most interesting outcome for labor economists involved labor supply. Strong advocates of the negative income tax program argued that the program would provide positive, or at least no negative, work incentives. Many economists, however, were skeptical, hypothesizing that the results would show some negative effect on work effort. Early experimental results discussed in Ross (1970), argued that work effort did not decline for the treatment groups. In fact, as Ross (1970, p. 568) indicates "there is, in fact, a slight indication that the participants' overall work effort increased during the initial test period."

Since this initial exploration, other scholars have re-examined the experimental design and data, coming to a less optimistic appraisal. An excellent elucidation is Ashenfelter (1990), who notes that because of attrition it is not actually possible to simply tabulate the results. In this sense, and from the experimenters point of view, the experiments were flawed in part because the design took little advantage of the inherent advantages of randomization. Of course, the ultimate policy test is whether the income maintenance programs increased work incentives relative to the existing welfare system, which as Moffitt (1981) notes at that time had large benefit-reduction rates that may have discouraged work. In certain cases, the new approach did outperform existing incentive schemes, in others it did not.

More importantly for our purposes, the New Jersey income maintenance experiment is generally considered to be the first large-scale social experiment conducted in the US, for which Ross is given credit (Greenberg et al., 1999; Greenberg and Shroder, 2004).[13]

[12] The negative income tax rate works as follows. Assume that John is randomly inserted into the 100% guaranteed income ($3300), 50% negative tax rate treatment. What this means is that when the policy binds, for each $1 that John's family earns on its own, they receive $0.50 less in federal benefits. Thus, if John's family earns $2000 in year one, they would receive $1000 less in program benefits, or $2300, resulting in a total income of $4300. In this case, if in any year John's family earns $6600 or more, program benefits are zero.

[13] We emphasize large scale because there were a handful of other social experiments—such as the Perry Preschool Project begun in 1962—that preceded the New Jersey Income Maintenance experiment (Greenberg et al., 1999), and that are still being evaluated today (Heckman et al., forthcoming). A prevalent type of social experimentation in recent years is the paired-audit experiments to identify and measure discrimination. These involve the use of "matched pairs" of individuals, who are made to look as much alike as possible apart from the protected characteristics. These pairs then confront the target subjects, which are employers, landlords, mortgage loan officers, or car salesmen. The majority of audit studies conducted to date have been in the fields of employment discrimination and housing discrimination (Riach and Rich, 2002).

The contribution of Ross, along with the excellent early summaries of the virtues of social experimentation (Orcutt and Orcutt, 1968), appears to have been instrumental in stimulating the explosion in social experiments in the ensuing decades.[14,15]

Such large-scale social experiments have continued in the US, and have included employment programs, electricity pricing, and housing allowances (see Hausman and Wise, 1985, for a review). While this early wave of social experiments tended to focus on testing new programs, more recent social experiments tended to be "black box" in the sense that packages of services and incentives were proffered, and the experiments were meant to test incremental changes to existing programs.[16] This generation of social experiments had an important influence on policy, contributing, for instance, to the passage of the Family Support Act of 1988, which overhauled the AFDC program. Indeed, as Manski and Garfinkel (1992) note, in Title II, Section 203, 102 Stat. 2380, the Act even made a specific recommendation on evaluation procedures: "a demonstration project conducted ... shall use experimental and control groups that are composed of a random sample of participants in the program."

Much like the experimental contributions of the agricultural literature of the 1920s and 1930s, the large-scale social experiments conducted in the twentieth century influenced the economics literature immensely. Since the initial income maintenance social experiment, there have been more than 235 known completed social experiments (Greenberg and Shroder, 2004), each exploring public policies in health, housing, welfare, and the like. The early social experiments were voluntary experiments typically

14 The original negative income tax experiment led to three other early experiments on income maintenance, which drew samples from rural areas of North Carolina and Iowa (1970-72); Seattle and Denver (1970-78); and Gary, Indiana (1971-74). These experiments went beyond studying urban husband-wife couples that were studied in the New Jersey income maintenance experiment. For instance, the North Carolina/Iowa study was conducted by the Institute of Research on Poverty to explore behavior among the rural poor. Only one and two parent black households were studied in the Gary, IN test. The Seattle-Denver study represented the most comprehensive, including blacks, Chicanos, and whites who had either one or two parents in the household. By and large, the evidence gathered in these studies reinforced the main result in the New Jersey study, but these new studies highlighted additional insights that were important for policy making, such as in differences between male and female labor force participation, unemployment duration, and welfare participation.

15 An early social experiment in Europe was the study of Intensified Employment Services in Eskilstuna, Sweden. In 1975, a small-town employment office received a personnel reinforcement for three months and split a group of 410 unemployed job seekers who had been registered at the office for at least three months into a treatment group ($n = 216$) and a control group ($n = 194$). The control group received normal service and used the services of the office for an average of 1.5 hours over the course of the experiment, while the treatment group used office services for an average of 7.5 hours, allowing office personnel to work more intensely on the individual problems of the treatment subjects. The findings were that the percent of workers with a job at the end of the experiment, unemployment spells during the experiment, and earnings were all favorably influenced by the employment services studied. A discussion of this study, as well as other European social experiments in labor market policy can be found in Bjorklund and Regner (1996) and the various Digests of Social Experiments due to Greenberg, and Shroder. Two of the more famous examples are the Norwegian Training Experiment (Raaum and Torp, 1993) and the Restart Programme in the United Kingdom (White and Lakey, 1992).

16 For example, whereas over 80% of social experiments from 1962-74 tested new programs, since 1983 only roughly 33% did so (Greenberg et al., 1999).

designed to measure basic behavioral relationships, or deep structural parameters, which could be used to evaluate an entire spectrum of social policies. Optimists even believed that the parameters could be used to evaluate policies that had not even been conducted. As Heckman (1992) notes, this was met with deep skepticism along economists and non-economists alike, and ambitions have since been much more modest.

As Manski and Garfinkel (1992) suggest, this second wave of social experiments had a methodological influence within academic circles, as it provided an arena for the 1980s debate between experimental advocates and those favoring structural econometrics using naturally-occurring data. Manski and Garfinkel (1992) provide an excellent resource that includes insights on the merits of the arguments on both sides, and discusses some of the important methodological issues. Highlighting some of the weaknesses of social experiments helps to clarify important distinctions we draw between social experiments and the generation of field experiments which has followed.

1.1.4. Potential shortcomings of social experiments

One potential problem arising in social experiments is "randomization bias", a situation wherein the experimental sample is different from the population of interest because of randomization. It is commonly known in the field of clinical drug trials that persuading patients to participate in randomized studies is much harder than persuading them to participate in non-randomized studies (Kramer and Shapiro, 1984). The same problem applies to social experiments, as evidenced by the difficulties that can be encountered when recruiting decentralized bureaucracies to administer the random treatment (Hotz, 1992).[17]

Doolittle and Traeger (1990) provide a description of the practical importance of randomization bias when describing their experience in implementing the Job Training Partnership Act. Indeed, in almost any social experiment related to job training programs, it is a concern that those most likely to benefit from the program select into the program. Moreover, as Harrison and List (2004) discuss, in social experiments, given the open nature of the political process, it is almost impossible to hide the experimental objective from the person implementing the experiment or the subject, opening up the possibility of such self-selection. As Heckman (1992) puts it, comparing social experiments to agricultural experiments: "plots of ground do not respond to anticipated treatments of fertilizer, nor can they excuse themselves from being treated."

To see this more formally, we follow the notation above and assume that $\tau_i = y_{i1} - y_{i0}$ is the treatment effect for individual i. Figure 1 shows a hypothetical density of τ_i in the population, a density assumed to have mean, τ^*. In this case, the parameter τ^* is equivalent to the average treatment effect; this is the treatment effect of interest if the analyst is pursuing an estimate of the average effect in this population.

[17] There is a growing body of evidence from laboratory settings on how individuals self-select into treatments when allowed to do so. See Lazear et al. (2009) for a recent such study, and the discussion in Charness and Kuhn (2011, 2010).

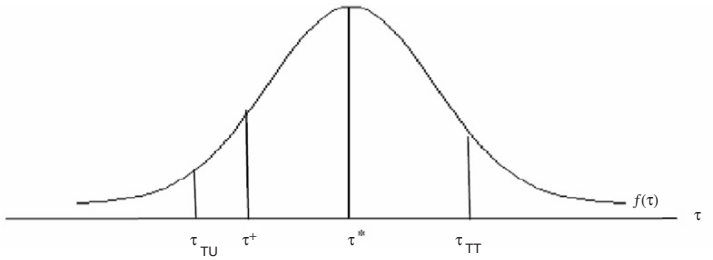


Figure 1 ***Simple illustration of the selection problem.***

The concern is that selection into the experiment is not random, but might occur with a probability related to τ. Using this notion to formulate the selection rule leads to positive selection: subjects with higher τ values are more likely to participate if offered. In Fig. 1, we denote the cutoff value of τ_i as τ^+: people above τ^+ participate, those below do not.

In this case, the treatment effect on the treated is what is measured in the social experiment: τ_{TT}. τ_{TT} is equal to $E(\tau_i|\tau_i > \tau^+)$, which represents the estimate of the treatment effect for those who select to participate. A lack of recognition of selection causes the analyst to mis-measure the treatment effect for the population of interest. Figure 1 also shows the treatment effect on the untreated, τ_{TU}. This τ_{TU} is equal to $E(\tau_i|\tau_i < \tau^+)$, which represents the unobserved estimate of the treatment effect for those who chose not to participate. Therefore, the population parameter of interest, τ^*, is a mixture of these two effects: $\tau^* = \Pr \times \tau_{TT} + (1 - \Pr) \times \tau_{TU}$, where Pr represents the probability of $\tau_i > \tau^+$. Even if one assumes that the population density of τ_i among participants is isomorphic to the population density of inferential interest, such selection frustrates proper inference. A related concern is whether the density of τ_i in the participant population exactly overlaps with the population of interest.

A second issue stems from Heckman (1992), Heckman and Smith (1995), and Manski (1995), who contend that participants in small-scale experiments may not be representative of individuals who would participate in ongoing, full-scale programs. Such non-representativeness of the experimental sample could occur because of a lack of information diffusion, the reluctance of some individuals to subject themselves to random assignment, or resource constraints in full-scale programs that result in program administrators restricting participants to people meeting certain criteria. As a result, making inference on how individuals would respond to the same intervention were they to be scaled up is not straightforward.

A third set of concerns stem from the supply side of those implementing the social experiment as it is scaled up. For example, the quality of those administering the intervention might be very different from the quality of personnel selected to take part in the original social experiment. Moreover, the ability of administrative agencies to closely

monitor those charged with the actual implementation of the program might also vary as programs are scaled-up. These concerns might also apply to field experiments unless they are explicitly designed to allow for such possibilities. In general, the role played by program implementers in determining program outcomes remains poorly understood and is a rich area for future study both for field experiments and researchers in general.

A fourth concern that arises in social experiments is attrition bias. Attrition bias refers to systematic differences between the treatment and control groups because of differential losses of participants. As Hausman and Wise (1979) note, a characteristic of social experiments is that individuals are surveyed before the experiment begins as well as during the experiment, which in many cases is several years. This within-person experimental design permits added power compared to a between-person experimental design—because of the importance of individual effects. But, there are potential problems, as they note (p. 455): "the inclusion of the time factor in the experiment raises a problem which does not exist in classical experiments—attrition. Some individuals decide that keeping the detailed records that the experiments require is not worth the payment, some move, some are inducted into the military."[18]

Beyond sampling and implementation shortcomings, social experiments also run the risk of generating misleading inference out of sample due to the increased scrutiny induced by the experiment. If experimental participants understand their behavior is being measured in terms of certain outcomes, some of them might attempt to succeed along these outcomes. Such effects have been deemed "John Henry" effects for the control sample because such participants work harder to show their worth when they realize that they are part of the control group. More broadly, some studies denote such effects as "Hawthorne" effects; if these Hawthorne effects do not operate equally on the treatment and control group, bias is induced.[19]

Another factor that might lead to incorrect inference in a social experiment is control group members seeking available substitutes for treatment. This is denoted "substitution bias" in the literature, a bias that can result in significant understatement of the treatment effect. Substitution bias can occur if a new program being tested experimentally absorbs resources that would otherwise be available to members of the control group or, instead, if as a result of serving some members of a target group, the new program frees up resources available under other programs that can now be used to better serve members of the

[18] Problems of attrition are well known and detailed discussions can be found in Hausman and Wise (1979) and the various chapters in Manski and Garfinkel (1992).

[19] Note that the development field experiments that have arisen recently often have to confront this issue directly when making inference from their studies—even though subjects might not know that they are randomized, a survey is used to measure the outcomes so repeated interactions are a certainty. One paper that attempts to quantify the effects is due to Gine et al. (2007). In a similar spirit, Muralidharan and Sundararaman (2007) present evidence from a randomized control trial on educational interventions in India. They also present evidence to distinguish the effects of the intervention from the mere effects of being part of an observational study *per se*.

control group. The practical importance of substitution bias is provided in Puma et al. (1990) and Heckman and Smith (1995).

Although these concerns, as well as others not discussed here, need always to be accounted for, social experiments continue to be an important and valuable tool for policy analysis, as evidenced by two recent and notable large scale undertakings: Moving To Opportunity (Katz et al., 2001) and *PROGRESA* (Schultz, 2004), as well as the more recent social experiments documented in Greenberg and Shroder (2004).

1.2. Field experiments

Following from the first two periods of experimentation discussed above, the third distinct period of field experimentation is the most recent surge of field experiments in economics. Harrison and List (2004), List (2006) and List and Reiley (2008) provide recent overviews of this literature. The increased use of this approach reflects a long running trend in labor economics, and applied microeconomics more generally, to identify causal effects. This is not surprising given that nearly all of the central research questions in labor economics are plagued by econometric concerns related to the simultaneous determination of individual decisions related to the accumulation of human capital, self-selection into labor markets and careers. Furthermore, many of the key variables that underlie behavior in labor markets—such as motivation or talent—are either simply unmeasured or measured with error in standard surveys.

Field experiments form the most recent addition to the wave of empirical strategies to identify causal effects that have entered mainstream empirical research in labor economics since the mid 1980s. For example, these are based on fixed effects, difference-in-differences, instrumental variables, regression discontinuities, and natural experiments. Comprehensive reviews of these developments are provided in Angrist and Krueger (1999). At the same time as these research strategies have developed, greater emphasis has been placed on econometric methods that are robust to functional form and distributional assumptions. These include the development of semi-parametric and non-parametric estimation techniques. Reviews of these developments are provided in Moffitt (1999).

We view the increased use of field experiments to have its origins in the last decade in part because of an acceleration of three long-standing trends in how applied economic research is conducted: (i) the increased use of research designs to uncover credible causal effects; (ii) the increased propensity to engage in primary data collection; and (iii) the formation of ever closer interactions with practitioners and policy makers more generally.

Similar to the experiments at the Hawthorne plant and social experiments, but unlike the first-generation agricultural studies, the most recent field experiments typically apply randomization to human subjects to obtain identification. In contrast to social experiments, however, recent field experiments strive to carry out this randomization on naturally occurring populations in naturally occurring settings, often without the

research subjects being aware that they are part of an experiment. As a consequence, these more recent studies tend to be carried out opportunistically, and on a smaller scale than social experiments.[20]

This current generation of field experiments oftentimes has more ambitious theoretical goals than social experiments, which largely aim to speak to policy makers and identify whether a package of measures leads to some desired change in outcomes. Modern field experiments in many cases are designed to test economic theory, collect facts useful for constructing a theory, and organize primary data collection to make measurements of key parameters, assuming a theory is correct. Field experiments can also help provide the necessary behavioral principles to permit sharper inference from laboratory or naturally-occurring data. Alternatively, field experiments can help to determine whether lab or field results should be reinterpreted, defined more narrowly than first believed, or are more general than the context in which they were measured. In other cases, field experiments might help to uncover the causes and underlying conditions necessary to produce data patterns observed in the lab or the field.

Since nature in most cases does not randomize agents into appropriate treatment and control groups, the task of the field experimental researcher is to develop markets, constructs, or experimental designs wherein subjects are randomized into treatments of interest. The researcher faces challenges different from those that arise either in conducting laboratory experiments or relying on naturally occurring variation. The field experimenter does not exert the same degree of control over real markets as the scientist does in the lab. Yet, unlike an empiricist who collects existing data, the field experimenter is in the data generating business, as opposed to solely engaging in data collection or evaluation. Consequently, conducting successful field experiments demands a different set of skills from the researcher: the ability to recognize opportunities for experimentation hidden amidst everyday phenomena, an understanding of experimental design and evaluation methods, knowledge of economic theory to motivate the research, and the interpersonal skills to manage what are often a complex set of relationships involving parties to an experiment.

1.2.1. What is a field experiment?

Harrison and List (2004) propose six factors that can be used to determine the field context of an experiment: the nature of the subject pool, the nature of the information that the subjects bring to the task, the nature of the commodity, the nature of the task or trading rules applied, the nature of the stakes, and the environment in which the

[20] In this sense, field experiments parallel the research approach that exploits "natural experiments" (Meyer, 1995; Rosenzweig and Wolpin, 2000; Angrist and Krueger, 2001), the difference being that in a field experiment the researcher actually controls the randomization herself, whereas in the natural experiment approach the researcher attempts to find sources of variation in existing data that are "as good as randomly assigned." In addition, the close involvement of the researcher from the outset allows for primary data collection to perhaps directly help shed light on the underlying mechanisms driving causal effects.

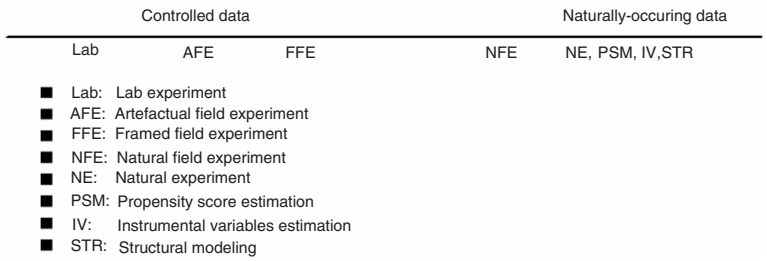


Figure 2 ***A field experiment bridge.***

subjects operate. Using these factors, they discuss a broad classification scheme that helps to organize one's thoughts about the factors that might be important when moving from non-experimental to experimental data.

They classify field experiments into three categories: artefactual, framed, and natural. Figure 2 shows how these three types of field experiments compare and contrast with laboratory experiments and approaches using naturally occurring non-experimental data. On the far left in Fig. 2 are laboratory experiments, which typically make use of randomization to identify a treatment effect of interest in the lab using a subject pool of students. In this Handbook, Charness and Kuhn (2011, 2010) discuss extant laboratory studies in the area of labor economics.

The other end of the spectrum in Fig. 2 includes empirical models that make necessary identification assumptions to identify treatment effects from naturally-occurring data. For example, identification in simple natural experiments results from a difference in difference regression model: $Y_{it} = X_{it}\beta + \tau T_{it} + \eta_{it}$, where i indexes the unit of observation, t indexes years, Y_{it} is the outcome, X_{it} is a vector of controls, T_{it} is a binary treatment variable equal to one if unit i is treated and zero otherwise, $\eta_{it} = \alpha_i + \lambda_t + \varepsilon_{it}$, and τ is measured by comparing the difference in outcomes before and after for the treated group with the before and after outcomes for the non treated group. A major identifying assumption in this case is that there are no time-varying, unit-specific shocks to the outcome variable that are correlated with T_{it}, and that selection into treatment is independent of a temporary individual specific effect.

Useful alternatives include the method of propensity score matching (PSM) developed in Rosenbaum and Rubin (1983). Again, if both states of the world were observable, the average treatment effect, τ, would equal $\overline{y}_1 - \overline{y}_0$. However, given that only y_1 or y_0 is observed for each observation, unless assignment into the treatment group is random, generally $\tau \neq \overline{y}_1 - \overline{y}_0$. The solution advocated by Rosenbaum and Rubin (1983) is to find a vector of covariates, Z, such that $y_1, y_0 \perp T|Z$, $\Pr(T = 1|Z) \in (0, 1)$, where $\perp$ denotes independence. This assumption is called the "conditional independence assumption" and intuitively means that given Z, the non-treated outcomes are what the treated outcomes would have been had they not

been treated. Or, likewise, that selection occurs only on observables. If this condition holds, then treatment assignment is said to be "strongly ignorable" (Rosenbaum and Rubin, 1983, p. 43). To estimate the average treatment effect (on the treated), only the weaker condition $E[y_0|T = 1, Z] = E[y_0|T = 0, Z] = E[y_0|Z]\Pr(T = 1|Z) \in (0, 1)$ is required. Thus, the treatment effect is given by $\tau = E[\overline{y}_1 - \overline{y}_0|Z]$, implying that conditional on Z, assignment to the treatment group mimics a randomized experiment.[21]

Other more popular methods of estimating treatment effects include the use of instrumental variables (Rosenzweig and Wolpin, 2000) and structural modeling. Assumptions of these approaches are well documented and are not discussed further here (Angrist and Krueger, 1999; Blundell and Costa-Dias, 2002). Between the two extremes in Fig. 2 are various types of field experiment. We now turn to a more patient discussion of these types.

1.2.2. A more detailed typology of field experiments

Following Harrison and List (2004), we summarize the key elements of each type of field experiment in Table 1. This also makes precise the differences between field and laboratory experiments.

Harrison and List (2004) argue that a first useful departure from laboratory experiments using student subjects is simply to use "non-standard" subjects, or experimental participants from the market of interest. In Table 1 and Fig. 2, these are denoted as "artefactual" field experiments. This type of field experiment represents a potentially useful type of exploration outside of traditional laboratory studies because it affords the researchers with the control of a standard lab experiment but with the realism of a subject pool that are the natural actors from the market of interest. In the past decade, artefactual field experiments have been used in financial applications (Alevy et al., 2007; Cipriani and Guarino, 2009), to test predictions of game theory (Levitt et al., 2009), and in applications associated with labor economics (Cooper et al., 1999).[22]

[21] Several aspects of the approach are not discussed in this discussion. For example, for these conditions to hold the appropriate conditioning set, Z, should be multi-dimensional. Second, upon estimation of the propensity score, a matching algorithm must be defined in order to estimate the missing counterfactual, y_0, for each treated observation. The average treatment effect on the treated (TT) is given by,

$$\tau_{TT} = E[E[y_1|T = 1, p(Z)] - E[y_0|T = 0, p(Z)]] = E[E[y_1 - y_0|p(Z)]],$$

where the outer expectation is over the distribution of $Z|T = 1$. These and other issues are discussed in List et al. (2003).

[22] Harrison and List (2004) discuss in detail whether student subjects exhibit different behaviors in laboratory environments that individuals drawn from other subject pools. A parallel trend in laboratory settings has been the use of "real-effort" experiments, as discussed in Charness and Kuhn (2011, 2010).

Table 1 A typology of field experiments.

	Non experimental	Natural experiments	Natural field experiments	Framed field experiment	Artefactual field experiment	Laboratory
In the field	×	×	×	×		
Real incentives	×	×	×	×		
Real task or information	×	×	×	×		
Aware of experiment				×	×	×
Appropriate people	×	×	×	×	×	
Researcher intervenes			×	×	×	×
Exogenous change		×				

Another example of the use of artefactual field experiments is to explain or predict non-experimental outcomes. An example of this usage is Barr and Serneels (2009), who correlate behavior in a trust game experiment with wage outcomes of employees of Ghanian manufacturing enterprises. They report that a one percent increase in reciprocity in these games is associated with a fifteen percent increase in wages. Another example is Attanasio et al. (2009) who combine household data on social networks with a field experiment conducted with the same households in Colombia, to investigate who pools risk with whom when risk pooling arrangements are not formally enforced. Combining non-experimental and experimental research methods in this way by conducting an artefactual field experiment among survey respondents provides an opportunity to study the interplay of risk attitudes, pre-existing networks, and risk-sharing.

Moving closer to how naturally-occurring data are generated, Harrison and List (2004) denote a "framed field experiment" as the same as an artefactual field experiment, except that it incorporates important elements of the context of the naturally occurring environment with respect to the commodity, task, stakes, and information set of the subjects. Yet, it is important to note that framed field experiments, like lab experiments and artefactual field experiments, are conducted in a manner that ensures subjects understand that they are taking part in an experiment, with their behavior subsequently recorded and scrutinized. Framed field experiments include the Hawthorne plant experiments, the social experiments of the twentieth century, and two related experimental approaches.

One related approach might be considered a cousin of social experiments: the collection of studies done in developing countries that use randomization to identify causal effects of interventions in settings where naturally-occurring data are limited. The primary motivation for such experiments is to inform public policy. These studies typically use experimental treatments more bluntly than the controlled treatments discussed above, in that the designs often randomly introduce a package of several interventions. On the other hand, this package of measures is directly linked to a menu of actual public policy alternatives. A few recent notable examples of this type of work are the studies such as Kremer et al. (2009) and Duflo et al. (2006).

Framed field experiments have also been done with a greater eye towards testing economic theory, for instance several framed field experiments of this genre have been published in the economics literature, ranging from further tests of auction theory (Lucking-Reiley, 1999; Katkar and Reiley, 2006), tests of the theory of private provision of public goods (Bohm, 1984; List, 2004a), tests that examine the relative predictive power of neoclassical theory versus prospect theory (List, 2003b, 2004b), tests that explore issues in cost/benefit analysis and preference elicitation (List, 2001, 2002a, 2003a; Lusk and Fox, 2003; Rozan et al., 2004; Ding et al., 2005)), tests that explore

competitive market theory in the field List, 2002b, 2004c; List and Price, 2005), and tests of information assimilation among professional financial traders (Alevy et al., 2007).[23]

Unlike social experiments, this type of framed field experiment does not need to worry about many of the shortcomings discussed above. For example, since subjects are unaware that the experiment is using randomization, any randomization bias should be eliminated. Also, these experiments tend to be short-lived and therefore attrition bias is not of major importance. Also, substitution bias should not be a primary concern in these types of studies. The cost of not having these concerns is that rarely can the long run effects of experimentally introduced interventions be assessed. This might limit therefore the appropriateness of such field experiments for questions in labor economics in which there are long time lags between when actions are made and outcomes realized.

As Levitt and List (2007a,b) discuss, the fact that subjects are in an environment in which they are keenly aware that their behavior is being monitored, recorded, and subsequently scrutinized, might also cause generalizability to be compromised. Decades of research within psychology highlight the power of the role obligations of being an experimental subject, the power of the experimenter herself, and the experimental situation (Orne, 1962). This leads to our final type of field experiment—"natural field experiments," which complete Table 1 and Fig. 2.

Natural field experiments are those experiments completed in cases where the environment is such that the subjects naturally undertake these tasks and where the subjects do not know that they are participants in an experiment. Therefore, they neither know that they are being randomized into treatment nor that their behavior is subsequently scrutinized. Such an exercise is important in that it represents an approach that combines the most attractive elements of the lab and naturally-occurring data: randomization and realism. In addition, it is difficult for people to respond to treatments they do not necessarily know are unusual, and of course they cannot excuse themselves from being treated. Hence, many of the limitations cited above are not an issue when making inference from data generated by natural field experiments. As we document in later sections, natural field experiments have already been used to answer a wide range of traditional research questions in labor economics.

1.2.3. Simple rules of thumb for experimentation

Scholars have produced a variety of rules of thumb to aid in experimental design. Following List et al. (2010), we provide a framework to think through these issues. Suppose that a single treatment T results in (conditional) outcomes Y_{i0} if $T = 0$, where $Y_{i0}|X_i \sim N(\mu_0, \sigma_0^2)$, and Y_{i1} if $T = 1$, where $Y_{i1}|X_i \sim N(\mu_1, \sigma_1^2)$. Since the experiment has not yet been conducted, the experimenter must form beliefs about the variances of outcomes across the treatment and control groups, which may, for example,

[23] Of course, this is just a select sampling of the work of this sort, for a more comprehensive list please see www.fieldexperiments.com.

come from theory, prior empirical evidence, or a pilot experiment. The experimenter also has to make a decision about the minimum detectable difference between mean control and treatment outcomes, $\mu_1 - \mu_0 = \delta$, that the experiment is meant to be able to detect. In essence, δ is the minimum average treatment effect, $\bar{\tau}$, that the experiment will be able to detect at a given significance level and power. Finally, we assume that the significance of the treatment effect will be determined using a t-test.

The first step in calculating optimal sample sizes requires specifying a null hypothesis and a specific alternative hypothesis. Typically, the null hypothesis is that there is no treatment effect, i.e. that the effect size is zero. The alternative hypothesis is that the effect size takes on a specific value (the minimum detectable effect size). The idea behind the choice of optimal sample sizes in this scenario is that the sample sizes have to be just large enough so that the experimenter: (i) does not falsely reject the null hypothesis that the population treatment and control outcomes are equal, i.e. commit a Type I error; and, (ii) does not falsely accept the null hypothesis when the actual difference is equal to δ, i.e. commit a Type II error. More formally, if the observations for control and treatment groups are independently drawn and $H_0 : \mu_0 = \mu_1$ and $H_1 : \mu_0 \neq \mu_1$, we need the difference in sample means $\bar{Y}_1 - \bar{Y}_0$ (which are of course not yet observed) to satisfy the following two conditions related to the probabilities of Type I and Type II errors.

First, the probability α of committing a Type I error in a two-sided test, i.e. a significance level of α, is given by,

$$\frac{\bar{Y}_1 - \bar{Y}_0}{\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}}} = t_{\alpha/2} \Rightarrow \bar{Y}_1 - \bar{Y}_0 = t_{\alpha/2}\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}}, \tag{1}$$

where σ_T^2 and n_T for $T \in \{0, 1\}$ are the conditional variance of the outcome and the sample size of the control and treatment groups. Second, the probability β of committing a Type II error, i.e. a power of $1 - \beta$, in a one-sided test, is given by,

$$\frac{\left(\bar{Y}_1 - \bar{Y}_0\right) - \delta}{\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}}} = -t_\beta \Rightarrow \bar{Y}_1 - \bar{Y}_0 = \delta - t_\beta\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}}. \tag{2}$$

Using (1) to eliminate $\bar{Y}_1 - \bar{Y}_0$ from (2) we obtain,

$$\delta = \left(t_{\alpha/2} + t_\beta\right)\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}}. \tag{3}$$

It can easily be shown that if $\sigma_0^2 = \sigma_1^2 = \sigma^2$, i.e. $\text{var}(\tau_i) = 0$, then the smallest sample sizes that solve this equality satisfy $n_0 = n_1 = n$ and then,

$$n_0^* = n_1^* = n^* = 2\left(t_{\alpha/2} + t_\beta\right)^2 \left(\frac{\sigma}{\delta}\right)^2. \tag{4}$$

If the variances of the outcomes are not equal this becomes,

$$N^* = \left(\frac{t_{\alpha/2} + t_\beta}{\delta}\right)^2 \left(\frac{\sigma_0^2}{\pi_0^*} + \frac{\sigma_1^2}{\pi_1^*}\right), \tag{5}$$
$$\pi_0^* = \frac{\sigma_0}{\sigma_0 + \sigma_1}, \quad \pi_1^* = \frac{\sigma_1}{\sigma_0 + \sigma_1},$$

where $N = n_0 + n_1$, $\pi_0 + \pi_1 = 1$, $\pi_0 = \frac{n_0}{n_0 + n_1}$.

If sample sizes are large enough so that the normal distribution is a good approximation for the t-distribution, then the above equations provide a closed form solution for the optimal sample sizes. If sample sizes are small, then n must be solved by using successive approximations. Optimal sample sizes increase proportionally with the variance of outcomes, non-linearly with the significance level and the power, and decrease proportionally with the square of the minimum detectable effect. The relative distribution of subjects across treatment and control is proportional to the standard deviation of the respective outcomes. This suggests that if the variance of outcomes under treatment and control are fairly similar—namely, in those cases when there are expected to be homogeneous treatment effects—there should not be a large loss in efficiency from assigning equal sample sizes to each.

In cases when the *outcome* variable is dichotomous, under the null hypothesis of no treatment effect, $\mu_0 = \mu_1$, one should always allocate subjects equally across treatments. Yet, if the null is of the form $\mu_1 = k\mu_0$, where $k > 0$, then the sample size arrangement is dictated by k in the same manner as in the continuous case. If the cost of sampling subjects differs across treatment and control groups, then the ratio of the sample sizes is inversely proportional to the square root of the relative costs. Interestingly, differences in sampling costs have exactly the same effect on relative sample sizes of treatment and control groups as differences in variances.

In those instances where the unit of randomization is different from the unit of observation, special considerations must be paid to the correlation in outcomes between units in the same treated cluster. Specifically, the number of observations required is multiplied by $1+(m-1)\rho$, where ρ is the intracluster correlation coefficient and m is the size of each cluster. The optimal size of each cluster increases with the ratio of the within to between cluster standard deviation, and decreases with the square root of the ratio of the cost of sampling a subject to the fixed cost of sampling from a new cluster. Since the

optimal sample size is independent of the available budget, the experimenter should first determine how many subjects to sample in each cluster and then sample from as many clusters as the budget permits (or until the optimal total sample size is achieved).[24]

A final class of results pertains to designs that include several levels of treatment, or more generally when the treatment variable itself is continuous, but we assume homogeneous treatment effects. The primary goal of the experimental design in this case is to simply maximize the variance of the treatment variable. For example, if the analyst is interested in estimating the effect of treatment and has strong priors that the treatment has a linear effect, then the sample should be equally divided on the endpoints of the feasible treatment range, with no intermediate points sampled. Maximizing the variance of the treatment variable under an assumed quadratic, cubic, quartic, etc., relationship produces unambiguous allocation rules as well: in the quadratic case, for instance, the analyst should place half of the sample equally distributed on the endpoints and the other half on the midpoint. More generally, optimal design requires that the number of treatment cells used should be equal to the highest polynomial order of the anticipated treatment effect, plus one.

1.2.4. Further considerations

In light of the differences between field experimentation and other empirical methods—lab experiments and using observational data—it is important to discuss some perceived differences and potential obstacles associated with this research agenda. One shortcoming of field experiments is the relative difficulty of replication vis-à-vis lab experiments.[25] As Fisher (1926) emphasized, replication is an important advantage of the experimental methodology. The ability of other researchers to reproduce quickly the experiment, and therefore test whether the results can be independently verified, not only serves to generate a deeper collection of comparable data but also provides incentives for the experimenter to collect and document data carefully.

There are at least three levels at which replication can operate. The first and most narrow of these involves taking the actual data generated by an experiment and reanalyzing the data to confirm the original findings. A second notion of replication is to run an experiment which follows a similar protocol to the first experiment to determine whether similar results can be generated using new subjects. The third and most general

[24] Relatedly, there is a recent but steadily expanding literature in statistics and economics on how experimental evidence on treatment effect heterogeneity may be used to maximize gains from social programs. One example is Bhattacharya and Dupas (2010) who study the problem of allocating a binary treatment among a target population based on observables, to maximize the mean social welfare arising from an eventual outcome distribution, when a budget constraint limits what fraction of the population can be treated.

[25] This is especially so if we compare field experiments to laboratory experiments that utilize student subject pools. Even by changing the subject pool slightly, as in artefactual field experiments, replicability becomes an issue as more still needs to be understood on the self-selection into experiments of such non-standard subjects (Charness and Kuhn, 2011, 2010).

conception of replication is to test the hypotheses of the original study using a new research design.

Lab experiments and many artefactual and framed field experiments lend themselves to replication in all three dimensions: it is relatively straightforward to reanalyze existing data, to run new experiments following existing protocols, and (with some imagination) to design new experiments testing the same hypotheses.

With natural field experiments, the first and third types of replication are easily done (i.e. reanalyzing the original data or designing new experiments), but the second type of replication (i.e. re-running the original experiment, but on a new pool of subjects) is more difficult. This difficulty arises because by their very nature, many field experiments are opportunistic and might be difficult to replicate because they require the cooperation of outside entities or practitioners, or detailed knowledge and the ability to manipulate a particular market.

Another consideration associated with field experiments relates to ethical guidelines (Dunford, 1990; Levitt and List, 2009). The third parties that field experimenters often need to work with can be concerned by the need to randomize units of observation into treatments. The benefits of such an approach need to be conveyed, as well as a practical sense of how to achieve this. For example, given resource constraints, practitioners are typically unable to roll out interventions to all intended recipients immediately. The field experimenter can intervene to randomly assign the order in which individuals are treated (or offered treatment), not whether they eventually receive the treatment or not.

With the onset of field experiments, new issues related to informed consent naturally arise. Ethical issues surrounding human experimentation is of utmost import. The topic of informed consent for human experimentation were recognized as early as the nineteenth century (Vollmann and Winau, 1996), but the principal document to provide guidelines on research ethics was the Nuremberg Code of 1947. The Code was a response to malfeasance of Nazi doctors, who performed immoral acts of experimentation during the Second World War. The major feature of the Code was that voluntary consent became a requirement in clinical research studies, where consent can be voluntary only if subjects: (i) are physically able to provide consent; (ii) are free from coercion; and, (iii) can comprehend the risks and benefits involved in the experiment.

What is right for medical trials need not be appropriate for the social sciences. To thoughtlessly adopt the Nuremberg Code whole cloth for field experiments without considering the implications would be misguided. In medical trials, it is sensible to have informed consent as the default because of the serious risk potential in most clinical studies. In contrast, the risks posed in some natural field experiments in economics are small or nonexistent, although such risks are almost certain to become more heterogeneous across field experiments as this research method becomes more prevalent. Hence while there might be valid arguments for making informed consent the exception, rather than the rule, in a field experimental context, it is true to say that covert

experimentation remains hotly debated in the literature. For more detailed discussions, the interested reader should see Dingwall (1980) and Punch (1985).

There are certain cases in which seeking informed consent directly interferes with the ability to conduct the research (Homan, 1991). For example, for years economists have been interested in measuring and detecting discrimination in the marketplace. Labor market field studies present perhaps the deepest amount of work in the area of discrimination. The work in this area can be parsed into two distinct categories, personal approaches and written applications.

Personal approaches include studies that have individuals either attend job interviews or apply for employment over the telephone. In these studies, the researcher matches two testers that are identical along all relevant employment characteristics except the comparative static of interest (e.g., race, gender, age), and after appropriate training the testers approach potential employers who have advertised a job opening. Researchers "train" the subjects simultaneously to ensure that their behavior and approach to the job interview are similar.

Under the written application approach, which can be traced to Jowell and Prescott-Clarke (1970), carefully prepared written job applications are sent to employers who have advertised vacancies. The usual approach is to choose advertisements in daily newspapers within some geographic area to test for discrimination. Akin to the personal approaches, great care is typically taken to ensure that the applications are similar across several dimensions except the variable of interest.

It strikes us as unusually difficult to explore whether, and to what extent, race or gender influence the jobs people receive, or the wages they secured, if one had to receive informed consent from the discriminating employer. For such purposes, it makes sense to consider executing a natural field experiment. This does not suggest that in the pursuit of science, moral principles should be ignored. Rather, in those cases Local Research Ethics Committees and Institutional Review Boards (IRBs) in the US serve an important role in weighing whether the research will inflict harm, gauging the extent to which the research benefits others, and determining whether experimental subjects selected into the environment on their own volition and are treated justly in the experiment.

1.2.5. Limits of field experiments

Clearly, labor economists rarely have the ability to randomize variables directly related to individual decisions such as educational attainment, the choice to migrate, the minimum wage faced, or retirement ages or benefits. This might in part reflect why some active research areas in labor economics have been relatively untouched by field experiments, as described in more detail below. However, field experiments allow the researcher scope to randomize key elements of the economic environment faced that determine such outcomes. For example in the context of educational attainment, it is plausible to design field experiments that create random variation over the monetary costs of acquiring education, information on the potential returns to education, knowledge of

the potential costs and benefits of education, or changes in the quality of inputs into the educational production function. Given the early and close involvement of researchers and the fact that primary data collection effort is part of a field experiment, there is always the potential to mitigate measurement error and omitted variables problems that are prevalent in labor economics (Angrist and Krueger, 1999).

Social experiments and field experiments are relatively easy for policy makers to understand. When designed around the evaluation of a particular policy or intervention, it is more straightforward to conduct a cost benefit analysis of the policy than would be possible through other empirical methods. As discussed before, a concern of using social experiments relates to sample attrition. While such attrition is less relevant in many field experiments, it is important to be clear that this often comes at the cost of field experiments evaluating relatively short run impacts of any given intervention. How outcomes evolve over time—in the absence of the close scrutiny of the experimenter, or how interventions should be scaled up to other units and other implementers, remain questions that field experimenters will have to always confront directly. Along these lines, we will showcase a number of field experiments in which researchers have combined random variation they have engineered to identify reduced form causal effects, with structural modeling to make out of sample predictions.

A second broad category of concerns for field experiments relate to sample selection. These can take a number of forms relating to the non-random selection of individuals, organizations, and interventions. At the individual level and in cases in which written consent is required, as for social and laboratory experiments, the self selection of individuals into the field experiment needs to be accounted for. Relatedly, the timing of decisions over who is potentially eligible to participate are critical, and potentially open to manipulation or renegotiation.

At the organizational level, there exists concerns related to whether we observe a non-random selection of organizations, or practitioners self-select to be subject to a field experiments. Similar concerns arise for social experiments often from political economy considerations.

Finally, at the intervention level, a concern is that practitioners, with whom field experimenters typically need to work, might only be willing to consider introducing interventions along dimensions they *a priori* expect to have beneficial effects. On the one hand this begs the question of why such practices have not been adopted already. On the other hand, one benefit of field experimentation might be that through closer ties between researchers and practitioners, the latter are prompted to think and learn about how they might change their behavior in privately optimal ways, and can be assured they will be able to provide concrete evidence of any potential benefits of such changes.

A third category of concerns relate to how unusual is the intervention. Although many parameters can be experimentally varied, it is important to focus on those parameters that would naturally vary across economic environments, and to calibrate

the magnitude of induced variations based on the range of parameter values actually observed in similar economic environments. Introducing unusual types of variation, or variations of implausible or unusual magnitude, or those that do not accord with theory, will be hard to make generalizations from and will not easily map back to an underlying theory. At the very least, care needs to be taken to separately identify whether responses to interventions reflect changes in equilibrium behavior that will persist in the long run, or agent's short run learning how to behave in new or unusual circumstances induced by the experimenter.

Fourth, there can sometimes be concerns that the third parties researchers collaborate with, might be under resource constraints that lead to the same set of implementers simultaneously or sequentially dealing with treated and control populations. Such implementation might lead to contamination effects and some of the other biases discussed above in relation to social experiments. This can lead to the use of within subject designs, where the researchers engineer an exogenously timed change to the economic environment, rather than between subject designs. The field experiments we discuss in later sections utilize both approaches.

Taken together, most of these concerns can be summarized as relating to the "external validity" of any field experiment—namely the ability to extrapolate meaningfully outside of the specific economic environment considered. This feature remains key to the worth of many field experiments. Field experiments almost inevitably face a trade-off between understanding the specifics of a given context and the generalizability of their findings. This trade-off can be eased by implementing a field experiment that considers the sources of heterogenous effects, or that combines reduced form estimates based on exogenous variation with structural modelling to predict responses to alternative interventions or to the same intervention in a slightly different economic environment.

Finally, it is worth reiterating that although primary data collection is a key element of field experimentation, this raises the costs of entry and might limit the number of experimenters relative to other purely lab based approaches. As will be apparent in the remainder of this chapter, there remain many issues in labor economics in which field experiments have yet to penetrate. In part these limits might be due to lack of opportunities, in some cases it might be because the activities under study are clandestine or illegal, although we will discuss carefully crafted field experiments to explore issues of racial discrimination for example. However, in some cases it is because the nature of the research question is simply not amenable to field experimentation. For example, questions relating to the design of labor market institutions are likely to remain outside the realm of field experimentation. In these and other cases, the controlled environment of the laboratory is the ideal starting point for economic research. Indeed, in this volume, Charness and Kuhn (2011, 2010) discuss the large laboratory-based literature on multiple aspects of the design of labor markets—such as market clearing mechanisms and contractual incompleteness. More generally, they discuss in detail the relative merits

of laboratory and field experiments. We share their view that no one research method dominates the other, and that in many scenarios using a combination of methods is likely to be more informative.

1.3. Research in labor economics

An enormous range of research questions are addressed by labor economists today. While the core issues studied by labor economists have always related to labor supply, labor demand, and the organization of labor markets, to focus our discussion, we limit attention to a select few topics. These reflect long-standing traditional areas of work in labor economics.[26]

First, since the seminal contributions of Gary Becker and Jacob Mincer, research in labor economics, particularly related to labor supply, has placed much emphasis on understanding individual decision making with regards to the accumulation of human capital. This emphasis has widened the traditional purview of labor economists to include all decision making processes that affect human capital accumulation. These decisions are as broad as those taken in the marriage market, within the household, and those on the formation of specific forms of human capital such as investments into crime. By emphasizing the role of individual decision making, subfields in labor related to the accumulation of human capital might be especially amenable to the use of field experiments.

Second, the empirical study of labor demand has been similarly revolutionized by the rapid increase in the availability of panel data on individuals, the personnel records of firms, and matched employer-employee data.[27] This has driven and fed back into research on various aspects of labor demand such as labor mobility, wage setting, rent sharing, and more generally, on the provision of incentives within organizations. This set of questions are again motivated by understanding the behavior of individuals and firms, there are rich possibilities to advance knowledge in related subfields through the use of carefully crafted field experiments. Field experiments offer the potential for researchers to lead data collection efforts.

To cover these broad areas, we loosely organize the discussion so as to follow an individual as they make important labor related decisions over their life cycle. Hence we discuss the role of field experiments in answering questions relating to early childhood interventions and the accumulation of human capital in the form of education and skills. We then consider research questions related to the demand and supply of labor, and labor market discrimination. We then move on to consider research questions related to

[26] More detailed discussions of how the study of labor economics has evolved over time can be found in Freeman (1987) and Taber and Weinberg (2008).

[27] To understand the magnitude of this change, we note that Stafford (1986) finds that among the 759 papers published in six leading journals between 1965 and 1983, virtually none was based on microdata with individual firms or establishments as the unit of analysis.

behavior within firms: how individuals are incentivized within firms, and other aspects of the employment relationship. Finally, we end with a brief discussion of the nascent literature on field experiments related to household decision making.

Table 2 shows the number of published papers in selected subfields of labor economics in the decade prior to the last volume of the *Handbook of Labor Economics* (1990-99), and over the last decade (2000-09). The table is based on all published papers in the leading general interest journals of *The American Economic Review*, *Econometrica*, *The Journal of Political Economy, The Quarterly Journal of Economics* and the *Review of Economic Studies*.[28] We use the *Journal of Economic Literature* classifications to place journal articles into one subfield within labor economics.[29]

Table 2 highlights a number of trends in published research in labor economics. First, the number of labor economics papers published in the top-tier general interest journals has not changed much over time. There were 278 published between 1990 and 1999, and 315 published between 2000 and 2009. Some of this increase probably reflects an increased numbers of papers in these journals as a whole, rather than changes in the relative importance of labor economics to economists. Examining the data by subfield, we do see changes in the composition of published papers in labor. There are large increases in the number of papers relating to: (i) education and the formation of human capital; (ii) firm behavior and personnel economics; (iii) household behavior; (iv) crime. Some of these increases reflect the wider available of data described above, such as personnel data from firms and matched employer-employee data sets, and primary data collected on households. Field experiments—an important component of which is primary data collection—are well placed to reinforce these trends. Indeed, below we discuss how field experiments have contributed to the first three of these areas in which there has been an increase in labor economics papers.

We observe a decline in papers on the organization of labor markets—an area in which not many field experiments have been conducted, in part because these questions are not well suited to field experimentation. Finally, the remaining subfields on the demand and supply of labor and on ageing and retirement remain relatively stable over the last two decades, and here field experiments remain scarce, but there might be particularly high returns from such research designs being utilized.

Second, the balance between theoretical and empirical work has remained relatively constant over the two decades. In both time periods, there have been approximately double the number of empirical as theoretical papers published in labor economics. We do not know whether for other areas of economics approximately a third of published papers are theoretical, but as will be emphasized throughout, labor economics has no shortage of theories that carefully designed field experiments can help determine the

[28] The numbers do not include papers and proceedings volumes.

[29] Earlier reviews of trends in published papers in labor economics include Stafford (1986), Manser (1999), and Moffitt (1999).

Table 2 Published research in labor economics by decade.

	Year of publication: 1990-99			Year of publication: 2000-09		
Subfield	*Theory*	*Empirics*	*Total*	*Theory*	*Empirics*	*Total*
Demand for education/formation of human capital	7	29	36	10	46	56
The demand and supply of labor	10	58	68	20	53	73
Organization of labor markets	28	59	87	20	48	68
Firm behaviour/personnel economics	26	23	49	33	28	61
Household economics	8	20	28	12	27	39
Aging and Retirement	0	6	6	0	6	6
Crime	2	2	4	1	11	12
Total	81	197	278	96	219	315

The table is based on all published papers in the leading general interest journals of the American Economic Review, Econometrica, the Journal of Political Economy, the Quarterly Journal of Economics, and the Review of Economic Studies. The numbers do not include papers and proceedings volumes. We use the Journal of Economic Literature classifications to place journal articles into one subfield within labor economics.

empirical relevance of. Within each subfield there are nearly always more empirical papers published than theoretical, with the exception of research into firm behavior and personnel economics, a pattern that holds across both decades. In other subfields, the ratios of theory to empirical papers vary considerably. Some areas such on the demand for education and formation of human capital have four to five times as many empirical papers, and the subfield of crime has been largely empirically driven.

1.3.1. How have labor economists used field experiments?

Table 3 presents evidence on the approach used by published papers in labor economics over the last decade.[30]

Three factors stand out. First, field experiments have been widely used in labor economics over the past decade, with there being 25 published papers using this research methodology in some form. For example, despite the surge in papers using laboratory experiments, over the last decade more papers published in the top-tier journals have employed field experiments. However, the number of empirical papers employing field experiments is still dwarfed by other empirical methodologies—there are 25 papers employing field experiments compared to 60 utilizing natural experiments, and 129 using non-experimental methods.

Second, the use of field experiments has thus far been concentrated to address research questions in a relatively small number of subfields in labor economics. Of the 25 published field experiments, three framed field experiments have been concerned with investments into education early in the life cycle, three natural field experiments have focused on the evaluation of specific labor market programs, and five natural field experiments have focused on incentives within firms.

In other subfields, such as on the determinants of wages and labor market discrimination, currently only one field experiment has been published, in contrast to four laboratory experiments. We view many research questions on discrimination in labor markets to be particularly amenable to study using field experiments. Hence this is one area in which field experiments have been relatively under supplied. Finally, the subfield of crime, which as documented in Table 2, has grown due almost exclusively to empirical papers, remains completely untouched by field experiments.

The third major fact to emerge from Table 3 is that there is a large supply of theory in labor economics, as previously noted in Table 2. Table 3 shows that this supply of theory is across all the subfields in labor economics. As we view carefully crafted field experiments to be able to potentially test between different theories, it would seem as if many areas of study of labor—across the life cycle from birth to retirement—are amenable to this method, and can give feedback on directions for future theoretical advancements.

[30] The total number of papers reported in Table 2 is not quite reflected in the totals recorded in Table 3. This is because in Table 3 we sometimes record a paper in more than one column if it utilizes a range of empirical techniques. For example, the total number of non-theory papers by subfield and method in Table 3 is greater than total non-theory papers found in Table 2 (224 > 219) because five papers used multiple methods and so were counted twice.

Table 3 Published papers 2000-9 by subfield and empirical method.

	Theoretical	Non experimental	Natural experiments	Natural field experiments	Framed field experiment	Artefactual field experiment	Laboratory
Demand for education/formation of human capital							
Early childhood interventions on human capital accumulation	2	3	3	1	0	0	0
(RCT) conditional cash transfer programs	1	2	1	1	1	0	0
Educational production function	3	6	9	0	3	0	0
Educational spillovers	1	3	2	0	0	0	0
Returns to education	3	8	3	0	0	0	0
The demand and supply of labor							
Wage and tax sensitivities	6	7	1	1	0	0	0
Determinants of wages/discrimination	12	28	4	1	0	0	4
Segmented labor markets	0	1	2	0	0	0	0
Demand for labor/skills	2	3	0	0	0	0	1

(*continued on next page*)

Table 3 (continued)

	Theoretical	Non experimental	Natural experiments	Natural field experiments	Framed field experiment	Artefactual field experiment	Laboratory
Organization of labor markets							
Unions, minimum wages and other labor market institutions	1	7	3	0	0	0	1
Labor market programs	0	1	3	3	0	0	0
Public sector labor markets	0	0	0	0	0	0	0
Occupational choice/intergenerational mobility/labor market segmentation	7	9	0	0	0	0	2
Immigration	0	6	2	0	0	0	0
Unemployment	12	13	3	0	0	0	0
Firm behavior/personnel economics							
Employee and executive incentives	16	4	4	5	0	0	2
The employment relationship/gift exchange	1	0	0	2	0	1	0
Peer effects	1	2	2	0	0	0	0
Workplace organization	15	4	0	2	0	0	0

Table 3 (continued)

	Theoretical	Non experimental	Natural experiments	Natural field experiments	Framed field experiment	Artefactual field experiment	Laboratory
Household economics							
Family size	1	3	2	0	0	0	0
Marital bargaining	2	0	0	0	1	0	0
The marriage market	6	4	3	1	0	0	0
Child labor	2	2	0	0	0	0	0
Household labor supply	0	2	1	0	0	0	0
Female participation	1	3	5	0	0	0	0
Retirement							
Decision to retirement	0	2	1	1	0	1	0
Health and retirement	0	1	0	0	0	0	0
Crime							
Crime	1	5	6	0	0	0	0
Total	96	129	60	18	5	2	10

The table is based on all published papers in the leading general interest journals of the American Economic Review, Econometrica, the Journal of Political Economy, the Quarterly Journal of Economics, and the Review of Economic Studies. The numbers do not include papers and proceedings volumes. We use the Journal of Economic Literature classifications to place journal articles into one subfield within labor economics.

To develop this point further, Table 4 provides a breakdown of how theory and evidence have been combined in labor economics, broken down by empirical method.

Two factors stand out. First, non-experimental papers are slightly more likely to use no theory than field experiments. Second, testing between more than one theory remains scarce, irrespective of the empirical approach. Although not all empirical papers should necessarily test theory, it is as important to establish facts on which future theory can be built. When testing between theories, it is important to both establish the power of these tests, to provide refutability or falsification checks, and to present evidence of the internal validity of the results. Natural field experiments might have a comparative advantage along such dimensions. Given such settings relate to real world behaviors, individuals are typically not restricted in how they respond to a change in their economic environment, which opens up the possibility of detecting behavior consistent with multiple theories.

Mirroring the discussion in Moffitt (1999), a second feature of on how best to use field experiments, that we aim to emphasize throughout, is the need to combine the use of field experiments with other research methodologies. For example, they might be combined with structural estimation, utilize a combination of evidence from the laboratory and the field, or draw inspiration from lab findings to establish plausible null and alternative hypotheses to be tested between.

Applying the full spectrum of approaches in trying to answer a single question can yield extra insights. A first example of such research relates to the importance of social preferences, which have been documented in numerous lab and field settings. To explore social preferences using a variety of approaches, List (2006) conducts artefactual, framed, and natural field experiments analyzing gift exchange. The games have buyers making price offers to sellers, and in return sellers select the quality level of the good provided to the buyer. Higher quality goods are costlier for sellers to produce than lower quality goods, but are more highly valued by buyers. The artefactual field experimental results mirror the typical findings with other subject pools: strong evidence for social preferences was observed through a positive price and quality relationship. Similarly constructed framed field experiments provide similar insights. Yet, when the environment is moved to the marketplace via a natural field experiment, where dealers are unaware that their behavior is being recorded as part of an experiment, little statistical relationship between price and quality emerges.

A second example comes from the series of field experiments presented in List (2004b)—from artefactual to framed to natural—in an actual marketplace to help distinguish between the major theories of discrimination: animus and statistical discrimination. Using data gathered from bilateral negotiations, he finds a strong tendency for minorities to receive initial and final offers that are inferior to those received by majorities in a natural field experiment. Yet, much like the vast empirical literature documenting discrimination that exists, these data in isolation cannot pinpoint the nature of discrimination. Under certain plausible scenarios, the results are consonant with at

Table 4 Testing theory by empirical method, published papers 2000-9.

	Research method	Theoretical	Non Experimental	Natural experiments	Natural field experiments	Framed field experiments	Artefactual field experiments	Laboratory
Method								
Pure theory	Theory	96	0	0	0	0	0	0
Theory and evidence	Test one theory	0	61	5	4	0	0	1
	Test between theories	0	10	6	5	1	1	5
Pure evidence	Pure empirics	0	58	49	9	4	1	4
		96	129	60	18	5	2	10

The table is based on all published papers in the leading general interest journals of the American Economic Review, Econometrica, the Journal of Political Economy, the Quarterly Journal of Economics, and the Review of Economic Studies. The numbers do not include papers and proceedings volumes. We use the Journal of Economic Literature classifications to place journal articles into one subfield within labor economics.

least three theories: (i) animus-based or taste-based discrimination, (ii) differences in bargaining ability, and (iii) statistical discrimination. By designing allocation, bargaining, and auction experiments, List (2004b) is able to construct an experiment wherein the various theories provide opposing predictions. The results across the field experimental domains consistently reveal that the observed discrimination is not due to animus or bargaining differences, but represents statistical discrimination.

1.4. Summary

We now move to describe, by various stages of the life cycle, how field experiments have been utilized in labor economics and the insights they have provided. Where appropriate, we discuss how these results have complemented or contradicted evidence using alternative research methods, and potential areas for future field experiments. We begin with the individual at birth and the accumulation of human capital before they enter the labor market. We then consider research questions related to the demand and supply of labor, and labor market discrimination. We then move on to consider research questions related to behavior within firms: how individuals are incentivized within firms, and other aspects of the employment relationship. Finally, we end with a brief discussion of the nascent literature on field experiments related to household decision making.

2. HUMAN CAPITAL

The literature associated with human capital acquisition prior to labor market entry is vast, and there is no room here to do it justice. As in the sections that follow, we therefore focus our discussion on a select few strands of this research and describe how field experiments can and have advanced knowledge within these strands. Even within this narrower branch of work, we are limited to focusing on select studies of inputs into the educational production function, where these inputs might be supplied by the school system, students, or their families.[31]

To see the issues, we follow Glewwe and Kremer's (2006) presentation of a framework for the education production function with the following reduced form representation,

$$S = f(C, H, Q, P), \tag{6}$$

$$A = h(C, H, Q, P), \tag{7}$$

where S is years of schooling, A is skills learned (achievement), C is a vector of child characteristics (including "innate ability"), H is a vector of household characteristics, Q is a vector of school and teacher characteristics (quality), and P is a vector of prices related to schooling. Q and P are both functions of education policies (EP) and local

[31] For related reviews of the literature, see the excellent work of Card (1999), and in this Handbook, the chapter by Fryer (2011).

community characteristics (L), which can be substituted into Eqs (6) and (7) to yield the following reduced form,

$$S = f(C, H, L, EP), \tag{8}$$
$$A = h(C, H, L, EP). \tag{9}$$

Similar to Mincerian human capital earnings functions, this framework estimates the partial equilibrium effects of educational inputs and policies, rather than general equilibrium effects that alter returns to education and thereby demand. Broadly, there are two approaches to estimating the production function.

The first focuses on measuring the effect of direct inputs, such as per pupil expenditure, class size, teacher quality and family background (i.e. estimating Eqs (6) and (7))). The second examines the effects of educational policies governing the structure of the school system (i.e., estimating Eqs (8) and (9)). In both cases, non-experimental and experimental estimates have shed insights into the relationships in the education production function (for a literature survey see Hanushek (1986)). To help place field experiments in this area in a wider context, we now turn to a non-exhaustive discussion of select work using both approaches, but not based on field experiments.

2.1. Measuring the effects of direct inputs

An early measurement study focusing on the effect of direct inputs is the report due to Coleman et al. (1966), who explored what fraction of the variation in student achievement could be explained by direct inputs. The Coleman report found only a weak association between school inputs and outputs. Subsequent regression based approaches largely replicated the findings in the Coleman report. Yet, one remarkably consistent result did emerge from these early studies: students situated in classrooms with a larger number of students outperformed children in smaller classes on standardized tests. This result is robust to inclusion of several conditioning variables, such as key demographic variables.

One aspect that this robust empirical finding highlights is the care that should be taken to ensure reverse causality and omitted variable bias do not frustrate proper inference. Given that the simple regression approach potentially suffers from biases due to endogeneity of policy placement, omitted variables, and measurement error—i.e., it is almost always the case that some unobserved element of the vectors C, H, Q, P or L will be correlated with both the outcome and the observed variables of interest—researchers have sought out other means to explore the parameters of the production function.

One such approach uses natural experiments. One neat example is the work of Angrist and Lavy (1999), who use legal rules to estimate the effect of class size on student performance. Assume legal limits on class size prevent the number of students in a

classroom from exceeding 25. Then consider a particular school that has cohorts ranging from 70–100. Thus, if a cohort includes 100 children, we would have four classrooms of size 25, whereas if the cohort includes 76 children, we end up with 4 classrooms with 19 children occupying each. Angrist and Lavy (1999) compare standardized test scores across students placed in different sized classrooms and find that a ten-student reduction raises standardized test scores by about 0.2 to 0.3 standard deviations.

As Keane (forthcoming) points out, this type of approach has similar drawbacks associated with the simple regression framework in the Coleman report. For instance, incoming cohort sizes might not be determined randomly because high performing schools attract more students. Likewise, cohort size might be affected by parents reacting to large class sizes by sending their kids elsewhere for schooling. Similar issues revolve around teacher assignment to small and large classrooms, which might not be randomly determined.

In this way, the Angrist and Lavy (1999) estimates should be viewed as a first step in understanding the importance of class size on student performance. The next step is to deepen our understanding by exploring the robustness of these results. One approach is to look for more observational data, another is to use randomization directly—similar to the accidental randomization of natural experiments, purposeful randomization can aid the scientific inquiry.

A central figure in using randomization in the area of education is William McCall, an education psychologist at Columbia University who, at odds with his more philosophical contemporaries, insisted on quantitative measures to test the validity of education programs. For his efforts, McCall is credited as an early proponent of using randomization rather than matching as a means to exclude rival hypotheses, and his work continues to influence the field experiments conducted in education today.[32]

A landmark social experiment measuring the effects of classroom size is the Tennessee STAR experiment. In this intervention, more than 10,000 students were randomly assigned to classes of different sizes from kindergarten through third grade. Similar to the social experiments discussed in the first Section, the STAR experiment had both attrition bias and selection problems in that some students changed from larger to smaller classrooms after the assignment had occurred. Nevertheless, even after taking these problems into account, Krueger (1999) put together a detailed analysis that suggests there are achievement gains from studying in smaller classes.

Combined, these two examples indicate that there very well might be a statistically meaningful relationship between class sizes and academic achievement, but the broader literature has not concluded that to be necessarily true. Scanning the entire set of

[32] Rockoff (2009) presents an overview of a substantial, but overlooked, body of field experiments class size that developed prior to World War II.

estimates from natural experiments and field experiments one is left with mixed evidence on the effects of class size at various tiers of the education system (Angrist and Lavy, 1999; Case and Deaton, 1999; Hoxby, 2000a; Kremer, 2003; Krueger, 2003; Hanushek, 2007; Bandiera et al., 2010, forthcoming).

Lazear (2001) theorizes that class size is dependent upon the behavior of students. As disruptive students are a detriment to the learning of their entire class, he proposes that the optimal class size is larger for better-behaved students. In his model, larger classes may be associated with higher student achievement, and may in part explain the mixed results in previous studies. This is one area where a natural field experiment might be able to help. One can envision that a test of Lazear's (2001) theory is not difficult if the researcher takes the data generation process in her own hands: designing experimental treatments that interact class size with student behavior would permit an estimation of parameters of interest for measures of both class size and peer inputs into the educational production function.

The results from this literature, more generally, make it clear how one could move forward with a research agenda based on field experimentation. For instance, are there critical non-linearities in the relationship between class sizes and academic performance, as suggested for university class sizes in Bandiera et al. (2010, forthcoming)? One might argue that the effects of smaller class sizes drop to zero at some critical threshold due to lost peer effects. What about the composition of classrooms? Even though the effects of peer composition are mixed (Hoxby, 2000b; Zimmerman, 2003; Angrist and Lang, 2004; Hoxby and Weingarth, 2006; Lavy et al., 2008; De Giorgi et al., 2009; Duflo et al., 2009), it might be the case that gender balance plays a key role in the classroom.

Even if we were to find strong evidence that class size matters for academic performance and answer the questions posed above, Eqs (6)–(9) highlight other features that we must be aware of before pushing such estimates too far. What is necessary is proper measurement of the estimates of the parameters of the production function, as well as an understanding of the decision rules of school administrators and parents. The next step is to deepen our understanding by exploring whether other more cost effective approaches to improve student achievement exist, say by understanding the optimal investment stream in students: at what age level are resources most effective in promoting academic achievement?

One line of work that addresses this question is the set of social experiments that explore achievement interventions before children enter school. Given that Fryer (2011) presents a lucid description of such interventions, we only briefly mention them here. The landmark social experiment in this area is the Perry Preschool program, which involved 64 students in Michigan who attended the Perry Preschool in 1962. Since then, dozens of other programs have arisen that explore what works with early childhood

intervention, including Head Start, the Abcedarian Project, Educare, Tulsa's universal pre-kindergarten program, and several others too numerous to list (see Fryer's Table 5).[33]

As Fryer (2011) notes, outcomes in these programs exhibit substantial variance. And, even in those cases that were met with great success, the achievement gains faded through time. Indeed, in many cases once school started the students in these programs gave back all academic gains (Currie and Thomas, 1995, 2000; Anderson, 2008). Another fact with the bulk of these programs is that they exhibit much homogeneity, mostly following from the general design in the Perry Preschool program. Much has been learned about early childhood development in the previous several decades, and this presents the field experimenter with a unique opportunity to make large impacts on children's lives. As Fryer (2011) notes, incorporating new insights from biology and developmental psychology represent opportunities for future research.

Such estimates cause us to pause and ask whether resource expenditures affect academic performance at all. In this spirit, there is a large literature that explores how direct school inputs, such as school expenditures, influence student performance. As a whole, the early literature found only a weak relationship between overall school expenditures and student achievement, primarily because resources tend to be allocated inefficiently (see Hanushek, 2006; Glewwe and Kremer, 2006 for a review of the recent literature). In response to these findings, a growing area of research uses both natural experiments and field experiments to examine a wide range of targeted investments in order to identify the effects and compare the cost-effectiveness of various interventions. For example, experiments have been carefully designed to identify the returns (in terms of schooling or achievement) to inputs such as school supplies, additional teachers, remedial education, or computer programs (Banerjee et al., 2001; Angrist and Lavy, 2002; Kremer et al., 2002; Glewwe et al., 2004, 2003; Banerjee et al., 2007).

2.2. Teacher quality

While the evidence on the effect of per pupil expenditure, class size, and peer composition is mixed, teacher quality has been found to be clearly important. Hanushek (2007) finds that the differences between schools can be attributed primarily to teacher quality differences. Little of this variation, however, can be explained by either teacher salaries or observable characteristics such as education and experience (Rivkin et al., 2005; Hanushek, 2006). Just as it is difficult to identify high quality teachers, little is

[33] There is evidence that the first five years of life are critical for lifelong development. Hence resource poor or unstimulating environments early in life are likely to detrimentally impact children's cognitive, motor, social-emotional development, and their health status (Grantham-McGregor et al., 1991; Heckman and Masterov, 2005; Engle et al., 2007; Grantham-McGregor et al., 2007). As adults, they are more likely to have high fertility rates and are less likely to provide adequate stimulation and resources for their own children, thus contributing to the intergenerational transmission of poverty and economic inequality (Sen, 1999). The current debate, to which social experiments have contributed, focuses on understanding the types of intervention that might be effective for the child and their families, and cost-effective from society's viewpoint.

known about how to improve teacher quality and performance. Given the evidence that education and professional development are largely ineffective, there is a growing interest in the use of performance-based incentives to improve teacher quality and effort.

The design and implementation of such incentives raises several areas of future study that observational data and field experiments can adequately fill, including: (i) what are the performance effects on the incentivized tasks and how can incentives be designed to cost-effectively maximize these effects; (ii) what are the effects on non-incentivized tasks, and how can incentives be designed to avoid diversion of effort in multitasking; (iii) how do teachers (of different quality) sort into different incentive and pay structures; and (iv) how does sorting affect general equilibrium teacher quality.

Evidence from non-experimental studies, natural experiments, and field experiments suggests that incentives can improve teacher performance (Lavy, 2002; Glewwe et al., 2003; Figlio and Kenny, 2007; Muralidharan and Sundararaman, 2007; Lavy, 2009; Duflo et al., 2009). Clearly, tighter links can be established between this literature and the larger labor literature on incentive design (Prendergast, 1999) to which, as discussed below, field experiments are also beginning to contribute.

More broadly, field experiments exploring mechanism design issues, such as comparing piece rate and tournament incentives, are rare. Also, these programs generally load incentives onto a single performance measure such as teacher attendance or student test scores, raising concerns that teachers might divert effort away from non-incentivized tasks (Holmstrom and Milgrom, 1991). Here, the evidence is mixed with some studies finding broad improvements in teacher effort (Duflo et al., 2009; Lavy, 2009) and others finding evidence of narrow efforts, such as teaching to the test, that divert effort from other tasks and do not improve long term student achievement (Kremer, 2006; Jacob, 2005).

Similarly, teacher sorting into incentive and pay structures is largely unexplored. Lazear (2001) applies his analysis of performance pay and productivity in a company (discussed in further detail below) to teacher incentives, suggesting that the effects of incentives on sorting could be comparable to effects on teacher effort. A well-designed field experiment could explore whether and how teacher sorting on incentives occurs. In general, field experiments that apply theories about incentive design, sorting and selection from other areas of labor could make a large contribution to the teacher incentives literature. Many of these issues arise in a later section when we discuss the role of field experiments in understanding behavior within firms.

Along with the school inputs, the primary inputs into the educational production function come from students and their families. A large literature models the effect of individual characteristics, family background and parental resources on schooling and achievement (Cameron and Heckman, 2001; Cameron and Taber, 2004). While it is impossible to randomly assign characteristics to individuals or to randomly assign children to families, quasi-experimental studies have exploited variation due to adoption in order

to separately identify genetic inputs ("nature") from parental inputs ("nurture") (Plug and Vijverberg, 2003). Other studies focus on potential barriers to individual investment in human capital production. These include high costs to education, perhaps due to credit constraints or high discount rates, and low marginal returns to education due, for example, to poor health or lack of human capital investment prior to entering school.

Estimates from non-experimental studies and natural experiments suggest that credit constraints are of limited importance in schooling decisions (Cameron and Taber, 2004; Stanley, 2003). However, estimates from natural experiments and field experiments of Conditional Cash Transfer programs (CCTs) find largely positive and significant effects, suggesting that (at least among the population targeted by CCTs) reducing present costs of education can affect schooling and human capital investment decisions. Non-experimental studies, natural experiments and field experiments have also found positive and significant effects from conditional cash transfer programs based on enrollment, attendance, and performance (Cornwell et al., 2006; Barrera-Osorio et al., 2008; Angrist and Lavy, 2009; Maxfield et al., 2003; Kremer et al., 2009). Few of these experiments, however, explore how conditional cash transfers can be most effectively designed.

Berry (2009) develops a model of household education production in which parents' ability to motivate their children is dampened by moral hazard. He then designs incentives to test several predictions of the model including the ability of parents to commit and the relative efficacy of incentives awarded to parents or to children based on the relative productivity of the two parties. Similarly, Levitt et al. (2010) implement a field experiment that compares both the incentive recipient (parent or student) and the incentive mechanism (piece rate or lottery). They also compare a year long broad-based incentive program that motivates sustained effort to an immediate one-time incentive aimed solely at increasing effort on a single standardized test. This design allows the authors to test a model of family investment, responsiveness to incentive mechanisms, and human capital returns from varying levels of effort. Both of these field experiments illustrate that researchers can design instruments that build on and test economic theory.

While conditional cash transfers aim to induce improvements in achievement by motivating greater effort and investment, a second strand of interventions attempts to directly improve abilities that can improve achievement. A growing of interest in this area focuses on investment in early childhood. Researchers argue that improving the abilities of young children can have long run returns on educational achievement, attainment and other outcomes such as employment, crime, fertility and health (Cunha and Heckman, 2009). Evidence from non-experimental studies, natural experiments and field experiments suggest that early education interventions can have significant effects on lifetime outcomes (Currie and Thomas, 1995; Currie, 2001; Garces et al., 2002; Behrman et al., 2004; Todd and Wolpin, 2006; Ludwig and Miller, 2007; Heckman et al., 2010).

Most of these studies require econometric techniques, such as matching, to correct for lack of valid randomization. And all of them are limited to identifying the effect of the intervention as a whole. They are not able to explore, for example, the relative importance of educational interventions compared to interventions that increase parental investments in early childhood. Given the evidence that early childhood is a key period of development and the relatively sparse body of empirical work, field experiments could address open questions related to: (i) the short and long run returns of the various inputs of the educational production function; (ii) to collect primary data and design field experiments to help decompose overall changes in outcomes from any given intervention into those arising from the behavioral responses of children, parents and teachers. Akin to the literature on public and private transfers to households (Albarran and Attanasio, 2003), this second strand of research can help shed light on whether altering some inputs leads to other inputs in the educational production function to be crowded in or out.

A final strand of the literature focuses on improving child health as a means of increasing school attendance rates. Estimates from natural experiments and field experiments find that health interventions have a positive and significant effect on school attendance (Bleakley, 2007; Bobonis et al., 2006; Miguel and Kremer, 2004). Miguel and Kremer (2004) expand beyond identification of individual returns to health interventions, modeling the positive externalities of deworming ignored in previous estimations. They use a field experiment randomized over schools to estimate positive externalities on the health and school attendance of untreated children in treated schools and schools neighboring treated schools. They also examine effects on test performance and estimate the health care and educational cost effectiveness of the program. As the authors argue, studies that ignore positive externalities in the comparison groups will underestimate the effect of the intervention by missing the external effects of deworming and underestimating the direct effect in comparison with an inflated baseline, biasing treatment effects towards zero. They point out that this identification problem is well recognized in the labor literature estimating the effects of job training programs on both participants and non-participants. The authors suggest an extension of their study that randomizes treatment at various levels such as within schools, across schools, and within clusters of schools.

2.3. Measuring the effects of policies governing the system

Recent studies of educational policy exploit natural experiments with randomized lotteries and variation in school district density to estimate the effects of school competition, school choice, school vouchers, school accountability and the presence of relatively autonomous public schools, such as charter schools (Clark, 2009; Cullen et al., 2006; Hoxby, 2000c; Jacob, 2004; Rouse, 1998; Angrist et al., 2002; Abdulkadiroglu et al., 2009). While proponents of expanding school choice argue that, as in other

markets, choice and competition will improve overall school quality and efficiency, the empirical studies find somewhat mixed evidence on these educational policies.[34]

For example, non-experimental studies, natural experiments and field experiments finding that vouchers improve educational achievement include Peterson et al. (2003), Krueger and Zhu (2004), Angrist et al. (2002, 2006). On the other hand, using randomized school lotteries, Cullen et al. (2006) find that school choice programs have little or no effect on academic achievement, and they suggest that this result may be due to parents making poor choices. Hastings and Weinstein (2008) explore this hypothesis using both a natural experiment and a natural field experiment to examine how reducing information costs affects parental choices. In the natural experiment, parents listed their preferences for schools within a district both before and after receiving information mandated by No Child Left Behind (NCLB). The natural field experiment randomized distribution of a simplified version of the NCLB information to parents who had also received NCLB information and to parents who had received no information.

This design allows the authors to measure the effect of each piece of information alone as well as their interaction. They find that information on school-level academic performance pushes parents to choose higher scoring schools (with no differences across the types of information received). Using IV estimation, they also argue that these choices lead to increased academic achievement.

Similarly, a growing body of research has begun to identify the right tail of the distribution of treatment effects among heterogeneous charter schools (Dobbie and Fryer, 2009; Hoxby and Muraka, 2009; Angrist et al., 2010). These studies rely on randomized lotteries in oversubscribed schools and can only identify the effect of a school (or school system) as a whole. They have reported suggestive evidence, however, on specific features that correlate with successful schools, such as longer days, longer school years, highly academic environments and so on. Field experiments could be used to complement this work by separately identifying the effects of charter school innovations, such as length of school day, school time, and general environmental conditions on the educational production function.

The field experiments discussed in this section highlight several important advantages of their usage for labor economists. For example, they can address biases in previous empirical estimates, including those from non-experimental studies. They are able to build in empirical and theoretical literature from several fields, such as education, health, and labor. Finally, they can be used to identify parameters beyond the direct return of an input into an individual educational production function and explore mechanism design issues.

[34] Several theoretical papers suggest that school vouchers will lead to overall welfare gains, increased stratification, and efficiency gains (Epple and Romano, 1998; Ferreyra, 2007; Nechyba, 2000; Rouse, 1998; Figlio and Rouse, 2006; Hsieh and Urquiola, 2006; Epple et al., 2006; Arcidiacono, 2005).

In the end, it is clear that empirical explorations into human capital acquisition prior to labor market entry is invaluable, and that there are several approaches that can be used in concert to learn more about the important parameters of interest. We argue that in this area field experiments can usefully add to the knowledge gained from naturally-occurring data, and the many low apples that are left to be picked give us great confidence that field experiments will only grow in importance in tackling particulars in the educational production function.

3. LABOR MARKET DISCRIMINATION

Philosophers as far removed as Arcesilaus, Heraclitus, and Plato have scribed of injustice and extolled upon the virtues of removing it for the betterment of society. Perhaps taking a lead from these scholars, social scientists have studied extensively gender, race and age based discrimination in the marketplace. In this section we explore the stage of the life cycle where individuals are entering the labor market. We focus mainly on discrimination in labor markets and how field experiments can lend insights into this important social issue.

We begin with a statistical overview of the data patterns in labor market outcomes across minority and majority agents. To make precise how field experiments might be carefully designed, we need to discuss theories for why such discrimination exists. The two major economic theories of discrimination that we discuss are: (i) certain populations having a general "distaste" for minorities (Becker, 1957) or a general "social custom" of discrimination (Akerlof, 1980); (ii) statistical discrimination (Arrow, 1972; Phelps, 1972), which is third-degree price discrimination as defined by Pigou (1920)—marketers using observable characteristics to make statistical inferences about productivity or reservation values of market agents.

Empirically testing for marketplace discrimination has taken two quite distinct paths: regression-based methods and field experiments. The former technique typically tests for a statistical relationship between an outcome measure, such as wage or price, and a group membership indicator. By and large, regression studies find evidence of discrimination against minorities in the marketplace.[35] Field experimental studies, which have arisen over the past 35 years, typically use matched pairs of transactors to test for discrimination. Due to the control that field studies offer the experimenter, they have become quite popular and have by now been carried out in at least ten countries (Riach and Rich, 2002). Across several heterogeneous labor markets, as well as product markets as diverse as home insurance and new car sales, field studies have made a strong case that systematic discrimination against minorities is prevalent in modern societies.

[35] A comprehensive summary of the regression-based literature on discrimination are contained in Altonji and Blank (1999) and Yinger (1998).

While regression-based empirical studies have served to provide an empirical foundation that indicates discrimination is prevalent in the marketplace, they have been less helpful in distinguishing the causes of discrimination. As Riach and Rich (2002) note, findings from field studies appear to be more consistent with the majority white populations having a general "distaste" for minorities in the sense of Becker (1957) or a general "social custom" of discrimination in line with Akerlof (1980); but statistical discrimination (Arrow, 1972; Phelps, 1972), or marketers using observable characteristics to make statistical inference about productivity or reservation values of market agents, for example, cannot be ruled out, *ex ante* or *ex post*.

Before one can even begin to discuss social policies to address discrimination, it is critical to understand the causes of the underlying preferential treatment that certain groups receive. As has been emphasized throughout, the potential for field experiments to be explicitly designed to test between theories, is a key advantage of this approach over other methodologies. In this section, we provide a framework for how field experiments can be used to advance our understanding of not only the extent of discrimination in the marketplace but also the nature of discrimination observed.

3.1. Data patterns in labor markets

As Altonji and Blank (1999) noted, researchers have observed labor market differences across race and gender lines for decades. Yet, the magnitude of market differences, and hence what a new generation of field experiments seek to explain, has changed substantially over time. For example, there was convergence in the black/white wage gap during the 1960s and early 1970s, but such convergence lost steam in the two decades afterwards. In addition, the Hispanic/white wage gap has risen among both males and females in the 1980s and 1990s. Of course, the world has not remained stagnant since the 1990s, and this section is meant to update the results in Altonji and Blank (1999).

Table 5 presents the labor outcomes of whites, blacks, and Hispanics by gender in 2009. Table 5 includes a set of labor market outcomes by race and gender that labor economists have studied for decades. The data are based on tabulations from the Current Population Survey (CPS) from May 2009. Row 2 of Table 5 indicates that white men earn 13% (21%) more than white women (black and Hispanic men) on an hourly basis. Black and Hispanic women earn less than minority men and majority women.

When one focuses on annual earnings, row 3 of Table 5, the differential between white men continues: they earn more than 20% higher wages than minority men. Yet, for women the racial difference becomes markedly higher—50% for white women to black and 30% for white women to Hispanic. The differentials remain when we focus on full-time employees—rows 7 and 8 of Table 5. In general Table 5 tells a story that has been told often before: white men earn more money for hours worked than other groups, and white women earn more than their female counterparts.

Table 5 Labor market data by race and gender.

	White males	Black males	Hispanic males	White males	Black females	Hispanic females
All workers (2009)						
(1) Share of all workers	0.355	0.054	0.086	0.315	0.061	0.059
(2) Hourly wage	$17	$14	$14	$15	$13	$12
	(10.3)	(7.4)	(6.9)	(8.8)	(6.8)	(6.7)
(3) Annual earnings	$64,642	$53,252	$50,573	$62,293	$41,533	$40,266
	(52182.5)	(48829.9)	(36977.0)	(55483.4)	(42769.5)	(39587.2)
(4) Weeks worked	48.01	49.29	48.13	47.32	48.03	46.67
	(9.8)	(8.5)	(9.4)	(10.6)	(9.9)	(11.9)
(5) Hours worked per week	40.2	39	38.3	35	36.9	34.9
(6) Share part time	0.132	0.139	0.132	0.279	0.188	0.267
Full-time-full year (2009)						
(7) Hourly wage	$19	$15	$14	$16	$14	$13
	(10.3)	(7.6)	(7.1)	(8.1)	(6.6)	(6.2)
(8) Annual earnings	$66,928	$56,855	$51,517	$61,948	$44,510	$52,756
	(48880.6)	(50287.3)	(36652.8)	(47603.9)	(44894.0)	(41008.6)
All persons						
(9) Share over employed	0.812	0.726	0.829	0.720	0.679	0.695
(10) Unemployment rate (Jan 2010)	11.40%	20.70%	14.00%	7.50%	14.20%	12.30%
(11) Employment rate (Jan 2010)	63.50%	51.50%	66.10%	54.30%	51.20%	50.00%

Standard deviations are in parentheses.
Source: Current population survey, May 2009.

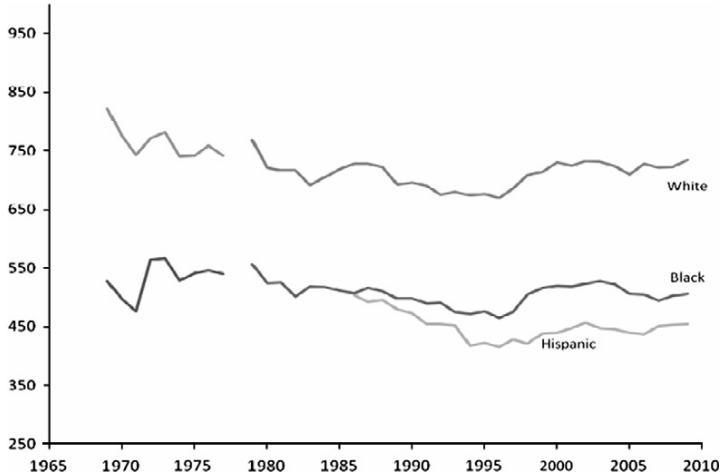


Figure 3 ***Median weekly earnings of male workers. Year 2000 dollars. (Source: Bureau of Labor Statistics.)***

Figures 3 and 4 complement these wage data by showing for each gender, the time series of annual median weekly income from 1969 to present for whites and blacks and 1986 to present for Hispanics.[36] These figures bring to light some interesting trends. Regardless of racial or ethnic group, wage rates for women continue to grow faster than for men. Within each gender, though, the 2000s did very little for racial or ethnic differentials. In fact, for both genders any local trend of convergence is reversed by the mid-2000s. In part this could be a function of the well documented rise in wage inequality during the second half of the 2000s.

Another important set of data points on Table 5 is the extent to which whites face lower unemployment rates. Figures 5 and 6 extend this information by showing, by gender, the time series of unemployment rates for whites, blacks, and Hispanics. One interesting element is the magnitude of unemployment changes for whites versus blacks and Hispanics. The mid-2000s saw no change to this trend, with the impact of recessions falling harder on blacks and Hispanics relative to whites. This trend does not seem to depend strongly on gender either, even though neither gender nor any racial or ethnic group seems to be immune from being hit by the 2009/10 recession.

Wages and unemployment rates are a function of labor force participation rates as well. Figure 7 shows the time series of labor force participation. The convergence in participation rates from the 70s through 90s continued into the 2000s, although

[36] Weekly earnings figures are taken from the Current Population Survey. They are for all employed people over age 25 that reported weekly earnings above zero. Data before 1979 is taken from the May supplement of the CPS. After 1979 data is taken from the CPS Annual Earnings File. Earnings from the May supplement for 1969-1972 were reported in ranges. The midpoint of each range was assumed to be the actual earnings for each individual.

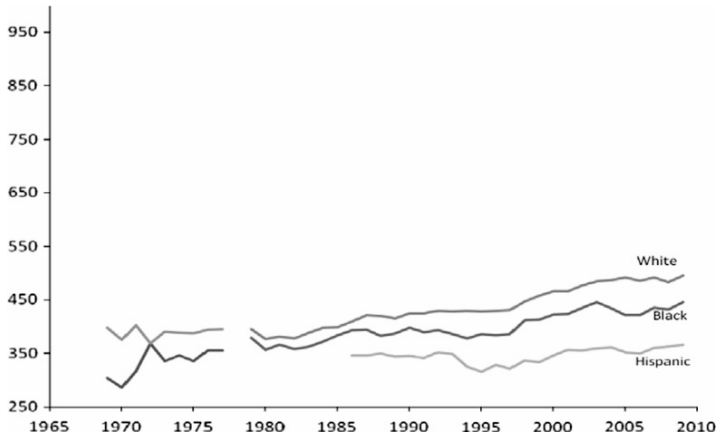


Figure 4 ***Median weekly earnings of female workers. Year 2000 dollars. (Source: Bureau of Labor Statistics.)***

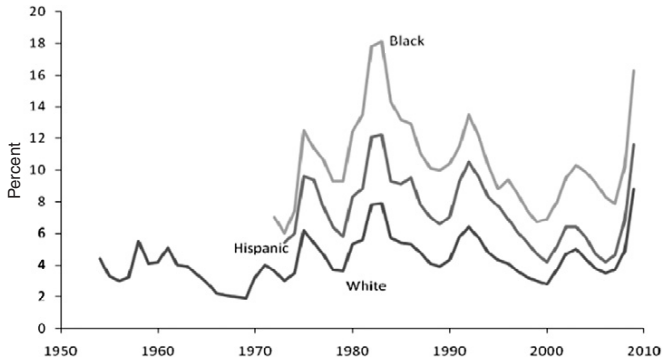


Figure 5 ***Male unemployment rates (annual averages) for men over 20. (Source: Bureau of Labor Statistics.)***

the pace of that convergence has slowed. Men of every race/ethnicity have dropped out of the labor force at a very slow rate while Hispanic and white females have increased participation. Interestingly, African American women have higher labor force participation than white women.

In considering the causes for these labor market disparities, economists have explored whether the workers themselves bring heterogeneous attributes to the workplace. To shed insights into this issue, we provide Table 6, which shows educational differences, family differences, and regional composition.

Rows 2 through 6 in Table 6 shows that whites obtain more years of education than blacks and Hispanics. Interestingly, white women are almost uniformly more educated then their ethnic/racial counterparts. This result is also reflected in the years

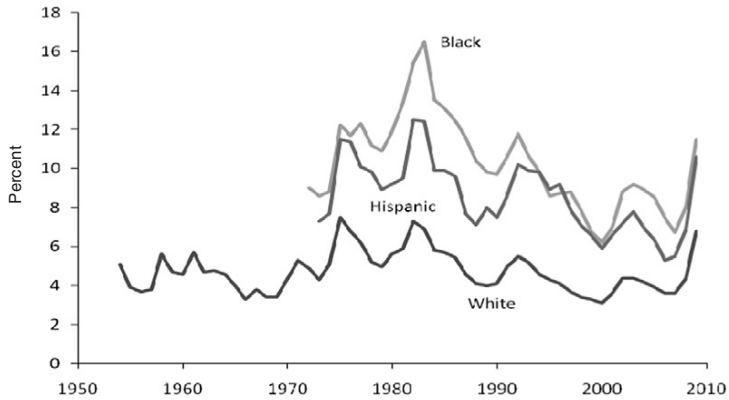


Figure 6 ***Female unemployment rates (annual averages) for women over 20. (Source: Bureau of Labor Statistics.)***

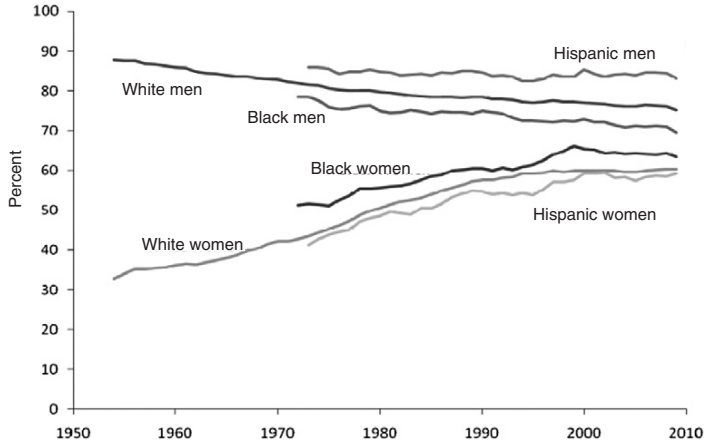


Figure 7 ***Labor force participation rates, 20 years and older. (Source: Bureau of Labor Statistics)***

of experience variable. Rows 8 through 10 in Table 6 give a sense of the different family choices (marriage and fertility) that are made by whites, blacks, and Hispanics—another important input to wages, especially for women. Row 8 shows that whites are more likely to be married (and perhaps enjoy the efficiencies of household trade), but row 10 shows that white women are likely to have fewer children to spend time caring for than black and Hispanic women. Rows 11 through 20 show the geographic breakdown of each race/ethnic group. Local labor market opportunities are surely influential for wages and in general, whites are from higher earning regions like New England and the Pacific.

Table 6 Personal characteristics by race and gender.

	All	White males	Black males	Hispanic males	White females	Black females	Hispanic females
(1) Share of all persons	1.00	0.32	0.05	0.07	0.35	0.07	0.07
Education							
(2) Less than high school	14.25%	9.88%	19.57%	38.58%	8.85%	16.96%	35.23%
(3) High school	30.92%	31.26%	36.86%	31.29%	31.11%	33.40%	28.94%
(4) Some post-HS training	19.66%	19.95%	20.21%	14.25%	20.76%	22.49%	16.77%
(5) College degree	26.11%	28.03%	18.55%	12.47%	29.58%	21.68%	15.82%
(6) More than college	9.07%	10.89%	4.81%	3.44%	9.70%	5.47%	3.24%
(7) Potential experience (Age-educ-5)	27.5 (18.3)	28.1 (17.8)	25.2 (17.5)	21.6 (16.3)	29.6 (18.9)	26.4 (18.3)	23.2 (17.3)
(8) Share married	0.545	0.597	0.409	0.52	0.557	0.3	0.497
(9) No. children age less than 6	0.148 (0.439)	0.128 (0.417)	0.101 (0.363)	0.199 (0.489)	0.139 (0.429)	0.16 (0.445)	0.26 (0.547)
(10) Total no. of children (age < 18)	0.809 (1.120)	0.707 (1.070)	0.804 (1.170)	1.18 (1.250)	0.747 (1.080)	0.985 (1.200)	1.244 (1.220)
(11) Share in SMSA*	0.789	0.751	0.888	0.912	0.752	0.891	0.918
Region							
(12) New England	11.2%	13.5%	4.2%	4.2%	13.7%	3.8%	5.0%
(13) Middle Atlantic	9.6%	9.3%	10.2%	9.7%	9.5%	11.5%	10.6%
(14) East-North Central	11.8%	13.1%	12.2%	6.0%	13.2%	12.5%	5.6%
(15) West-North Central	11.6%	14.1%	5.1%	4.1%	14.0%	4.5%	3.5%
(16) South Atlantic	17.9%	16.2%	38.7%	14.5%	16.5%	39.3%	13.9%
(17) East-South Central	5.0%	5.2%	9.4%	1.4%	5.3%	9.2%	0.9%
(18) West-South Central	8.3%	6.6%	10.6%	17.4%	6.6%	11.0%	17.9%
(19) Mountain	10.2%	10.8%	3.1%	14.4%	10.4%	2.7%	14.3%
(20) Pacific	14.5%	11.1%	6.5%	28.1%	10.8%	5.5%	28.4%

Source: Current population survey, March 2009. Standard deviations in parenthesis.

* Defined as residing in SMSA with at least one million inhabitants.

Overall, these data are in line with Altonji and Blank (1999), who find large educational differences among these groups, with race and ethnicity mattering much more than gender. Of course, what these education differences represent is difficult to parse. On the one hand, they might be mostly due to different preferences. Alternatively, they might reflect behavior of agents who expect to face discrimination later on in the labor market—referred to as "pre-market" discrimination. As Altonji and Blank (1999) note, there is evidence that some minorities have been denied market opportunities, perhaps leading to less than efficient levels of schooling investment.

While the labor market outcomes disparities observed in Table 5 and Figs 3–7 might represent differences mainly due to these individual investment choices, perhaps investment varies because of preferences, comparative advantage, and the like. For example, another hypothesis put forth is that such outcomes are at least partly due to discrimination in the labor market. The remainder of this section briefly discusses theories of discrimination and attempts to test these theories, contrasting regression-based approaches to field experiments.

3.2. Theories of discrimination

We follow the literature and define labor market discrimination as a situation in which persons who provide labor market services and who are equally productive in a physical or material sense are treated unequally in a way that is related to an observable characteristic such as race, ethnicity, or gender. By "unequal" we mean these persons receive different wages or face different demands for their services at a given wage.

We consider two main economic models: entrepreneurs are willing to forego profits to cater to their "taste" for discrimination, as first proposed by Becker (1957). The second model is "statistical" discrimination: in an effort to maximize profits, firm owners discriminate based on a set of observables because they have imperfect information. This could be as simple as employers having imperfect information on the relative skills or productivity of minority versus majority agents. The models in both of these literatures are deep, and rich with good intuition. We do not have the space to do them justice, but strive simply to provide a sketch of each model to give the reader a sense of how one can test between them. We urge the reader to see Altonji and Blank (1999) for a more detailed presentation of these models and their implications.[37]

[37] As far as the law is concerned, both types of discrimination—taste based and statistical—are illegal. For example, in credit markets, the Equal Credit Opportunity Act (Sec. 701, as amended in March 1976) states that it "shall be unlawful for any creditor to discriminate against any applicant, with respect to any aspect of the credit transaction. . . on the basis of race, color, religion, national origin, sex or martial status, or age. . . ". The law implies that while it is allowed to differentiate among customers based on characteristics of the customer (e.g., credit history) or the product that are linked to the expected return of the transaction, it is illegal to use the customer's membership in a group to distinguish among customers. In other words, firms should make decisions about the customer as if they had no information regarding the customer's race, sex, etc. This, for example, is true regardless of whether race is or is not a good proxy for risk factors in the credit market (Ladd, 1998).

3.2.1. Taste-based discrimination

In his doctoral dissertation, Becker (1957) modeled prejudice or bigotry as a "taste" for discrimination among employers. Becker modeled employers as maximizing a utility function that is the sum of profits plus a disutility term from employing minorities,

$$U = PF(N_{NM} + N_M) - W_{NM}N_{NM} - W_M N_M - dN_M, \tag{10}$$

where P is product price, F is the production function, which takes on two arguments: the number of employees that are non-minority N_{NM} and minority N_M. The second term is the wage bill and the final term is the disutility from employing minorities, dN_M. For prejudiced employers, the marginal cost of employment of a minority worker is $W_M N_M + dN_M$. Accordingly, d is the "coefficient of discrimination," or the level of distaste of the employer for employing a minority worker. The higher d, the more likely the employer will hire non-minority workers, even if they are less productive than minority workers.

The Becker model then shows that the wage premium for non-minority workers is determined by the preferences of the least prejudiced employer who hires minority workers. Several extensions to this model have been proposed in the literature, including the possibility that d is a function of the job type, wage level, or the extent of segregation in the labor market. For example, Coate and Loury (1993) develop a model that restricts all employers to have identical preferences, but makes d a factor only when the employer hires minority workers for skilled jobs; an important consideration in the model then becomes the ratio of minority and non-minority people working in skilled jobs.

A logical conclusion of many of the studies in this area is that with certain assumptions—in many cases free entry, constant returns to scale, segmenting, etc.—in the long run non-discriminating employers will increase to the point that it is no longer necessary for minority workers to work for prejudiced employers, eliminating any wage discrepancies between minority and non-minority workers. This is a testable implication.

3.2.2. Statistical discrimination

Arrow (1972) and Phelps (1972) discuss discrimination that is consistent with the notion of profit-maximization, or Pigou's (1920) "third-degree price discrimination." In this class of model, in their pursuit of the most profitable transactions, marketers use observable characteristics to make statistical inference about reservation values of market agents. The underlying premise implicit in this line of work is that employers have incomplete information and use observables to guide their behavior. For example, if they believe that women might be more likely to take time out of the labor force, employers with high adjustment costs might avoid those expected to have higher attrition rates. Firms then have an incentive to use gender to "statistically discriminate" among workers if gender is correlated with attrition.

Of course, employers can discriminate along second moments of observable distributions too. Sobel and Takahashi (1983) develop a model along such lines, and their model is reconsidered in List and Livingstone (2010), which we closely follow here. In this framework, employers look at second moments and use prior beliefs about the productivity of group members to influence hiring and wage outcomes.

In the case where workers approach employers in an effort to sell their labor services, the employer proposes the wage (price) in each period. The worker can accept or reject the offer. If the offer is rejected, the employer makes another offer. If the offer is rejected in the terminal period, no exchange occurs. To keep the analysis simple and without losing focus on the critical incentives, we consider a two-period model. The results can all be extended to an n-period model.

Consider the situation where the employer's reservation value is public information and denoted v_b, where $v_b \in [0, 1]$, and the employer knows only the distribution from which the worker's reservation valuation is drawn. An employer confronts a potential worker, who has reservation value v_s, which is drawn from a distribution $F(v)$ on support $[0, 1]$. It is assumed that this c.d.f. is continuously differentiable, and that the resulting p.d.f. $f(v)$ is positive for all $v_s \in [0, 1]$. The employer discounts future payoffs by $q, q \in [0, 1]$. Further assume that the worker discounts future payoffs at the rate of $p, p \in [0, 1]$. p and q can be thought of as the costs of bargaining and are known by both players.

The bargaining process proceeds as follows: the employer proposes a price (wage) to the worker in period 1. The worker can accept or reject the offer. If the offer is rejected, the employer proposes a new price. It is assumed that the new proposal must be a wage (price) that is no lower than the original offer. The worker can accept or reject this proposition. If it is rejected, the game ends and no transaction occurs.

Following a no-commitment equilibrium, the employer is assumed to make the period 1 offer at the beginning of period 1, and subsequently chooses the period 2 offer using the information gained from the worker's rejection of the period 1 offer. Let x_1 be the employer's offer in period 1, and let x_2 be the employer's offer in period 2. Also, define, for period $i = 1, 2$,

$$S_i = \begin{cases} 0 & \text{if } i = 0 \\ \dfrac{(x_1 - px_2)}{(1-p)} & \text{if } i = 1 \\ x_2 & \text{if } i = 2. \end{cases} \tag{11}$$

A worker whose reservation value is v_s will prefer accepting in period i to accepting in period $i + 1$ if $x_1 - v_s > p(x_{i+1} - v_s)$, or if $v_s < S_i$. A worker's most preferred time to accept is period i if $S_{i-1} < v < S_i$, so $F(S_i) - F(S_{i-1})$ is the employer's *ex ante* probability of hiring in the ith period and $(v_b - x_i)[F(S_i) - F(S_{i-1})]$ is the employer's *ex ante* undiscounted expected profit in period i.

The employer's maximization problem can be stated in terms of his choice of the period 2 offer, x_2, and of S_1, which implies a choice of x_1, since $S_1 = \frac{(x_1 - px_2)}{(1-p)}$. The employer's optimal strategies are found via backwards induction, starting with his period 2 decision. If the period 1 offer is rejected, then the employer knows $v_s \geq S$. The employer chooses an offer x_2 to maximize his expected profits. Let $\pi(S)$ be this maximum value,

$$\pi(S) = \max_{x_2} \frac{(v_b - x_2)[F(x_2) - F(S)]}{1 - F(S)}. \tag{12}$$

Let $x_2(S)$ be the unique value of x_2 that solves (12). The first order condition of this problem, which implicitly defines $x_2(S)$, implies that,

$$(v_b - x_2(S))f(x_2(S)) = F(x_2(S)) - F(S). \tag{13}$$

Since the offer $x_2(S)$ must be less than the employer's valuation v_b, the left-hand side of (13) is positive, so the right-hand side must also be positive. For this to be the case, it must be true that,

$$v_b > x_2(S) > S. \tag{14}$$

In other words, in equilibrium, the second period offer must be greater than the first period offer, and both offers must be less than v_b.

The no-commitment equilibrium is fully characterized by $x_2(S)$ and a first period price, $\widehat{x}_1$, that solves,

$$\max_x (v_b - x)[F(S(x))] + q\pi(S(x))[1 - F(S(x))], \tag{15}$$

subject to,

$$S(x) = \begin{cases} S, & \text{where } S = \dfrac{x - px_2(S)}{1 - p} \text{ if such as } S \in [0, 1] \text{ exists} \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

Substituting in the constraint and the definition of $\pi(S)$, the problem becomes,

$$\begin{aligned} \max_{0 \leq S \leq 1} \; & v_b F(S) - (1 - p)SF(S) \\ & - px_2(S)F(S) + q(v_b - x_2(S))[F(x_2(S)) - F(S)]. \end{aligned} \tag{17}$$

If $\widehat{S}$ solves (17), then $\widehat{x}_1 = (1-p)\widehat{S} + px_2(\widehat{S})$ and $\widehat{x}_2 = x_2(\widehat{S})$ are the no-commitment equilibrium offers. The first-order condition of (17) implies,

$$(1-q)v_b f(S) - (1-p)[F(S) - Sf(S)] + (q-p)x_2(S)f(S) + px_2'(S)F(S) = 0. \quad (18)$$

3.2.3. Optimal employer behavior

Within this framework one can analyze differences in how an employer will behave when he confronts members of the various groups. To obtain insights on the impact of changes in the variance of the worker's reservation value on both offers that the employer may make, we consider a simple example where the worker's value is drawn from a uniform distribution. There are two groups of potential workers. Members of group 1 draw their reservation value from a uniform distribution with lower bound a_1 and upper bound b_1. Members of group 2 draw their value from a uniform distribution with lower bound a_2 and upper bound b_2. Assume $a_1 > a_2$ and $b_1 < b_2$, so the variance of group 2's distribution is larger than the variance of group 1's distribution. Without loss of generality, further assume that the bounds are such that the distributions have equivalent means. Now, consider the employer's equilibrium offers, when confronting a worker who is a member of group i, $i = 1, 2$.

Solving through backwards induction, the employer first calculates his period 2 offer, as in (12). Substituting in the uniform distribution and simplifying, the problem becomes,

$$\max_{x_2}(v_b - x)\left(\frac{x-a}{b-S}\right) - (v_b - x)\left(\frac{S-a}{b-S}\right). \quad (19)$$

The solution of this problem is,

$$\widehat{x}_2 = x_2(S) = \frac{S + v_b}{2}. \quad (20)$$

Hence, the period 2 decision is a function of S, which is chosen in period 1. The employer's period 1 problem is to solve (17). Substituting in the distribution and the solution $x_2(S)$, the first order condition is given by,

$$\left(1 - \frac{1}{2}p\right)2S - \left(1 - \frac{1}{2}p\right)a + \frac{1}{2}q(S - v_b) = 0, \quad (21)$$

which implies the solution,

$$\widehat{S} = \frac{(1 - \frac{p}{2})a_i + \frac{1}{2}q v_b}{2 - p + \frac{q}{2}}, \quad (22)$$

making the optimal period 2 offer,

$$\widehat{x_2} = \frac{1}{4 - 2p + q}\left[\left(1 + \frac{q}{2}\right)v_b + \left(1 - \frac{p}{2}\right)a_i\right], \tag{23}$$

and the optimal period 1 offer,

$$\widehat{x_1} = (1-p)\frac{(1-\frac{p}{2})a_i + \frac{1}{2}q\,v_b}{2 - p + \frac{q}{2}} + p\frac{1}{4 - 2p + q}\left[\left(1 + \frac{q}{2}\right)v_b + \left(1 - \frac{p}{2}\right)a_i\right]. \tag{24}$$

Note that the optimal offers $\widehat{x_1}$ and $\widehat{x_2}$ are both increasing in a_i, and therefore decreasing in the variance of reservation values. Group 2's reservation value is drawn from a distribution with a larger variance, hence $a_1 > a_2$. In this example, then, the analysis shows that when the employer believes he is dealing with a member of a group whose reservation value is widely distributed (group 2), he will offer to hire at a lower wage than he would if the worker were a member of a group with a lower variance (group 1). This is true despite the fact that the first moments of the distributions are identical. This prediction provides one means to test the statistical discrimination model against the taste-based discrimination model. We return to this notion below.

3.3. Empirical tests

Scholars have concerned themselves primarily with the question "is there discrimination in market X?" and much less time has been spent on answering the question "why do firms discriminate?" As economists interested in public policy, however, we should be interested in not only the extent of discrimination but also the source of discrimination. Conditional on the existence of discrimination, it is imperative to understand the source of discrimination, since one cannot begin to craft social policies to address discrimination if its underlying causes are ill-understood. We now turn to an overview of a select set of studies that measure discrimination.

3.3.1. Observational data

One of the most important means to empirically test for marketplace discrimination in labor markets is to use regression-based methods. The focus using this approach has ranged from measuring labor force participation to modeling wage determination. Within the line of work that explores wages, the overarching theme is to decompose wage differentials, using an Oaxaca decomposition, between groups into what can be

explained by observables and what cannot be explained by observables. More specifically, consider a simple model that makes wages for minorities as follows,

$$W_M = \beta_M X_M + e_M, \tag{25}$$

and wages for non-minority agents as,

$$W_{NM} = \beta_{NM} X_{NM} + e_{NM}, \tag{26}$$

where W represent wages, X is a vector of individual specific observables that affect wages, and e is a classical error term. The wage difference between minority and non-minority agents can be computed by differencing these equations as follows,

$$\begin{aligned} W_M - W_{NM} &= [\beta_M X_M + e_M] - [\beta_{NM} X_{NM} + e_{NM}] \\ &= [X_M - X_{NM}]\beta_M + [\beta_M - \beta_{NM}]X_{NM}. \end{aligned}$$

The first term in the right most expression, $[X_M - X_{NM}]\beta_M$, is the component of the wage difference that is explained: it arises because of differences in the average characteristics of group members, such as region of residence, experience, or education level. The second term, $[\beta_M - \beta_{NM}]X_{NM}$, is the part of the wage difference that is not explained by the regression model—the differences in the response coefficients of the regression, or the rate of return differences across minorities and non-minorities. This last term encompasses differences in wages due to differences in the returns to similar characteristics between groups. For example, returns to education may differ across minorities and non-minorities. The fraction of wage difference due to this second term is typically called the "share" of wage differences due to discrimination.

Before discussing some of the general results from various regression-based approaches, it is important to qualify the results. First, the approach of assuming that the entire second component, $[\beta_M - \beta_{NM}]X_{NM}$, is due solely to discrimination is likely not correct. For example, for this to be true the wage equation must be well specified. If omitted variable bias exists, then the response coefficients will be biased. Second, this equation captures only discrimination in the labor market as measured today. That is, even if no discrimination is found to exist in such a model in today's wages, that does not imply discrimination is unimportant. For example, if women are constantly denied market opportunities for skilled jobs, they might not invest optimally to obtain such positions. In the literature, such under investment is denoted as market discrimination before, or "pre" market discrimination. Clearly, it is difficult to parse the effects of years past with the current effects of discrimination, and this should be kept in mind when interpreting the empirical results below—both those from the regression based model as well as from field experiments.

The regression models can be applied to the data discussed above from the Current Population Survey (CPS) from May 2009. Yet, given that Altonji and Blank (1999) summarize a series of regression results from such wage equations that do not differ markedly from ours, we simply restate the main results. First, we find white men receive significantly higher wages than black men, even after controlling for education, job experience, region of residence, and occupation. Following the letter of the model, this is evidence of discrimination in the data.

Second, even after controlling for key factors, Hispanic men and minority female workers have lower wages than their non-minority counterparts. Once again, if one sticks to the interpretation of the model, this is suggestive evidence that discrimination exists between these groups. One should highlight, however, that there are certain difficulties in using CPS data for such an exercise—such as the problem of not having individual ability measures, such as cognitive and non-cognitive abilities. Altonji and Blank (1999) extend the CPS results by modeling data from the National Longitudinal Survey of Youth (NSLY). In general, their results with NSLY data confirm that an improved specification reduces the unexplained effects for blacks and for women.

While this line of work is suggestive that discrimination exists in the labor market, due to productivity unobservables the nature of discrimination is not discernible without rather strong assumptions: are minority men receiving lower wages because of tastes or because of statistical discrimination?

Some headway has been made in these regards recently in several clever studies. One such study is the ingenuous paper of Goldin and Rouse (2000), who use audition notes from a series of auditions among national orchestras in order to determine whether or not blind auditions—those in which musicians auditioned behind a screen—help women relatively more than men. The authors use a panel data set and identify discrimination by the change in hiring practices toward blind auditions that occurred in the 1970s. Goldin and Rouse study the actual audition records obtained from orchestra personnel managers and orchestra archives from eight major symphony orchestras from the late 1950's to 1995. These records contain lists of everyone auditioning (first and last name) with notation around the names of those who advance. There are three rounds of auditions considered: preliminary, semifinal, and final. The gender of the participants is determined by their name (96% of the records are distinctly masculine or feminine).

Eighty-four percent of all preliminary rounds were blind, seventy-eight percent of all semifinal rounds were blind, and seventeen percent of all final rounds were blind. In addition, the authors have personnel rosters that describe final assignments (members of the orchestra). There is variation in hiring practices over time, so that within one orchestra, the same audition may be blind or non-blind over time and across categories (preliminary, semifinal, and final). In addition, since success is rare, the same musician sometimes auditions more than once.

In the authors' data, 42 percent of individuals competed in more than one round and 24 percent competed in more than one audition. Including musician fixed-effects, the authors identify the effect of a screen to hide gender from those individuals who auditioned both with and without a screen. Without this "ability" control (individual fixed-effects) the data suggests that women are worse off with blind auditions. However, controlling for individual fixed effects, the authors find that for women who make it to the finals, a blind audition increases their likelihood of winning by 33 percentage points.

In their main specification, the authors find that women are significantly less likely to advance from semifinals when auditions are blind, but significantly more likely to advance from preliminary auditions and final auditions when they audition behind a screen. Turning to the final outcome space—what is the effect of the screen on the hiring of women?—the authors estimate that though they are unable to obtain a statistically significant effect (since the likelihood of winning an audition is less than three percent), women are five percentage points more likely to be hired than men when auditions are completely blind and there is no semifinal round. There is no difference between the likelihood that women are hired relative to men when there is a semifinal round and auditions are blind.

Ultimately, the effects discussed give pause to reported "traditional" orchestra practices. In particular, "a strong presumption exists that discrimination has limited the employment of female musicians." Before the implementation of blind auditions, committees were instituted to overthrow the biased hiring practices of conductors (who reportedly hired select males from a small set of well known instructors). However, sex-based biases seemed to dominate hiring, even in the face of "democratization." As the authors demonstrate, the institution of blind hiring significantly increased the success rate of women in most auditions.

However, it is difficult for the authors to parse whether or not the discrimination is taste-based or statistical. The authors note that an orchestra is a team, which requires constant improvement and study together. In this sense, female-specific absences—maternity leave—can impact the quality of the orchestra significantly and may motivate statistical discrimination against women. Using their data, the authors note that the average female musician took 0.067 leaves of absence per year, compared to the average males' 0.061 leaves. The length of leave was negligibly different between genders. These statistics imply that taste-based discrimination, assuming no performance differences between hired males and females, are at least in part the cause of the discrimination against female musicians. Again, without the strong assumption that conditional on being hired, women and men of the same audition caliber perform indistinguishably in their careers, it is difficult for this innovative work to parse the type of discrimination observed.

A second clever piece of work based on the regression approach is due to Altonji and Pierret (2001), who create a model that generates strict predictions on the effect of

race on wages over time under a hypothesis of statistical discrimination based on race by employers. Notably, they conclude that if firms do not statistically discriminate based on race (if they follow the law), but race is negatively related to productivity, then: (i) the race gap will widen with experience, and, (ii) adding a favorable variable that the hiring firm cannot observe will reduce the race difference in the experience profile. The authors find that the data satisfy these predictions: the race gap widens with experience and the addition of a "skill" variable reduces the race gap in experience slopes. Thus, the authors conclude that employers "do not make full use of race as information."

Fundamentally, the authors' model studies the differential effect on wages of "easy to observe" s variables and "hard to observe" z variables that predict worker productivity. While s variables such as schooling should have a smaller and smaller effect on wage over time, since an employer's experience with the worker reveals far more important predictors of productivity, those variables that are difficult to observe such as skill have a relatively larger effect on wages as time goes on. This implies that the authors can identify whether or not the easily observable characteristic of race is acting as an s variable, or if employers are ignoring it. If employers are ignoring race, but race remains negatively correlated with productivity, then race acts as a z variable, appearing more important—more predictive of wage—over time. Again, the authors find support for the latter case.

The authors estimate their model using NLSY 1979 data—a panel study of men and women aged 14-21 in 1978 that have been surveyed annually since 1979. The data on white and black men with eight or more years of education forms the basis of their empirical analysis. The authors use AFQT (the Armed Forces Qualification Test) scores as a variable that employers do not observe, but that predicts productivity. In addition, the authors control for the first job held by all subjects in order to ensure that their results are not driven by the effect that a high AFQT may have on a worker's access to jobs in which skill is observed, rather than "dead-end jobs" where skill is never observed. Because the authors control for secular shifts in the wage structure, their identification of the interactions between time and observable (s) characteristics and unobservable or ignored (z) characteristics comes from variation across age cohorts.

The authors find that a one standard deviation shift in AFQT rises from having no effect on wages when experience is zero to increasing log wages by 0.0692 when experience is 10. This supports the result that employers learn about productivity. The coefficient on education interacted with experience declines from 0.0005 to −0.0269 when the variable AFQT*experience is added. With an intercept of 0.0832 with the addition, we can conclude that the effect of an extra year of education declines from 0.829 to 0.0595 over ten years. This suggests that employers statistically discriminate on the basis of education because they have limited information about labor market entrants. In short, the effect of easy-to-observe variables like education dwindles as hard-to-observe variables like ability become more available—as time goes on and the employer becomes more familiar with the quality of the worker. The authors find similar effects with their

other hard-to-observe variables that correlate with productivity such as father's education and sibling wage rate: as experience increases, these variables become more and more predictive of higher wages (though the effect of father's education is never significant).

The main analysis is on whether or not employers statistically discriminate based on race. If firms use race as information—that is, as easily-observable predictors of performance similar to education—then the effect of race over time on wages should decline as hard-to-observe variables like skill (predicted by the AFQT) become more transparent over time. If firms ignore race, however, the initial (experience $= 0$) race gap should be small, and should widen with experience if race is negatively related to productivity. Also, when race is ignored (a z variable) adding another z variable like AFQT$\star$experience will reduce the race gap in experience slopes. The authors note that the effect of a "black" dummy will not necessarily be zero even if firms do not statistically discriminate on the basis of race, since race may be correlated with legally usable information available to the employer but not to the econometrician.

Empirical analysis shows that the effect of adding AFQT$\star$experience decreases the race gap in experience slopes (from -0.1500 to -0.0861); this is the opposite of what we would expect if employers fully used race as a predictor of performance (as they do with schooling—recall, the addition of AFQT$\star$experience increases the amount by which the impact of education changes over time). Using another prediction of their model, that the effect of learning on the s variables will equal the effect of learning on the z variables times the relationship between the s and z variables—that there are spillover effects from learning—the authors are able to reject race as an s variable but not able to reject race as a z variable.

A few points are of note. First, if the quantity of training is influenced by the employer's beliefs about a worker's productivity, effects of training cannot be separated from the effects of statistical discrimination with learning. In addition, if taste-based discrimination becomes more prevalent at higher level positions, a widening of the race gap based on experience may be a reflection of increased taste-based discrimination rather than employer learning. Finally, the authors model the effect of statistical discrimination on wages, but not on the extended hiring decision. Based on these considerations, the authors note that any of their results on race-based discrimination should be interpreted cautiously.

To summarize, Altonji and Pierret test for statistical discrimination in a very reasonable way: they argue that if firms statistically discriminate, an observable characteristic such as race will be very important in predicting wages early in the employment history—before productivity is well observed—but becomes less important in predicting wages as time goes on and the worker accumulates experience. In the data the opposite is true, suggesting that under the assumption that the model is well specified firms attempt to ignore race in their hiring decisions, but that race is correlated with productivity (which is revealed) and so it becomes more and more predictive of wages as time goes on.

A third innovative regression-based study is due to Charles and Guryan (2008), who use state-level variation in historical wage and survey data to empirically test the impacts of discrimination on the labor market, focusing on taste for discrimination. The main theoretical result from Becker's work explored by Charles and Guryan is the assertion that black workers are hired by the least prejudiced employers in the market due to sorting in the labor market. Furthermore, they examine whether racial wage gaps are determined by the prejudice of a marginal employer, not the average. This sorting mechanism provides Charles and Guryan with two empirical regularities to verify Becker's work: (i) the level of prejudice observed by the employers displaying large amounts of prejudice (in the upper tail of a distribution of prejudice) should not impact wages; (ii) holding prejudice constant, wages should be lower with more blacks in the labor market.

Although they do not target the question of taste versus statistical based discrimination directly, they do include a variable for the skill difference between blacks and whites in regressions run as robustness checks. This and other robustness checks do not alter the main results which find support for Becker's theory of marginal prejudice affecting wages: marginal and low percentile prejudice levels negatively impact the black white wage gap while higher percentile and average prejudice levels have no impact; also the percent of the population that is black has a negative impact on the wage gap.

Charles and Guryan (2008) begin by empirically motivating the relationship between the black-white wage gap and prejudice by displaying the correlation between wage data from the CPS and white survey responses to questions concerning racial sentiments from the General Social Survey (GSS). After displaying the positive wage gap to prejudice relationship, Charles and Guryan review the theoretical findings to clarify the hypotheses of interest and then discuss the data. The data being used for prejudice is a non-uniform (the same questions are not asked every year) nationally representative survey with state-level data from 1972-2004. The survey questions used in this analysis are those from white responders and are vetted to reflect prejudice as much as possible (for example a question on whether "the government was obligated to help blacks" was not used due to the possible response aimed at the government). The survey responses were used to formulate a prejudice index relative to the responses given in 1977 and a prejudice distribution and the data on prejudice is combined with CPS May monthly supplement from 1977 and 1978 and CPS Merged Outgoing Rotation Group (MORG) for analysis.

The empirical results come from a hedonic wage regression. The regressions are run at the state level under the assumptions that employment markets are at the state level and interstate moves are costly. Because the prejudice measure they have is at the state level, Charles and Guryan take an additional step to allow for more reasonable standard errors than ones that would come from a full regression with observations at the individual level. This additional step comprises removing the prejudice index but including a black dummy variable for each state (state-black dummy interaction) in the first stage wage hedonic, and then using the coefficient from this interaction term as the

dependent variable in a second stage regression which includes the prejudice index. Five main measures of the prejudice index are analyzed: average, marginal, 10th percentile, median and 90th percentile. The marginal level of prejudice is calculated as the "pth percentile of the prejudice distribution, where p is the percentage of the state workforce that is black" and the prejudice distribution is calculated from the GSS data. Additionally, the fraction of the population that is black in the state is included in the second stage.

The second stage regression results all support Becker's theory. The first result of a negative impact on the black-white relative wages (negative means lower wages for blacks) attributed to the average level of prejudice is not significant and becomes positive when the marginal level of prejudice is included. The impact of the marginal prejudice measure is always negative and significant. This is also the case for the coefficient on the measurement of the fraction of the state population that is black (always negative and significant). These first results are taken as indication that the average prejudice measures fail to explain the wage gap, while the marginal and fraction of black have the assumed relationship from Becker's work.

The additional prejudice measurements: 10th percentile, median and 90th percentile, provide further support for Becker's theory. When included together in a regression, both with and without the percent of the state's population that is black, the 10th percentile is the only variable of significance (it is negative). This result is taken as further support of Becker's theory because of the indication that higher measurements of prejudice do not affect the wage gap (note that when the proportion of the state's population that is black is included, the 10th percentile increases in both absolute magnitude and significance).

Various robustness checks are completed such as the inclusion of variables to indicate skill as mentioned above. Two skill measures are used: (i) separate reading and math variables which measure the difference between black and white test scores at the state level from a National Assessment of Educational Progress-Long Term Trend (NAEP-LTT) test, and, (ii) black-white relative school quality measures used and Card and Krueger (1992) (for which they reduce the sample to just southern states). In both cases the results are similar to when the skill proxies are not included. Although this identification strategy does not disentangle the impact of taste and statistical based discrimination, the inclusion of skill level measures does suggest that this is taste-based discrimination under the assumption that the skill measures accurately reflect the difference in work-place abilities between races and that these differences in abilities are known by the employers. In a best case scenario, identifying statistical discrimination would require some measure of employee productivity by race and employment.

Further robustness checks investigate other possible endogeneity issues. An instrument of the proportion black in the state workforce in 1920 was used to account for possible endogeneity issues with the percent of the state's current population that is black. No difference in results was found. Finally, Charles and Guryan (2008) include a measure from the National Education Longitudinal Survey of 1988 (NELS) to account

for the fraction of co-workers that were of the same race. The results again supported Becker's theory that market sorting results in blacks being more segregated towards lower prejudiced employers: the wage gap is larger when the co-workers are more mixed when accounting for racial prejudice and the black proportion of the population.

The overall result is best restated directly from the last paragraph in the paper: "Our various results suggest that racial prejudice among whites accounts for as much as one-fourth of the gap in wages between blacks and whites... a present discounted loss in annual earnings for blacks between $34,000 and $115,000, depending on the intensity of the prejudice of the marginal white in their states."

Similar to the above studies, making an assumption on the regression specification, allows Charles and Guryan (2008) to begin to parse the type of discrimination observed. As such, as all of these incredibly insightful studies illustrate, one can go a long way in detecting discrimination, and its sources, but pinpointing exactly the extent that taste based and statistical discrimination is the underlying motive, is only possible with additional assumptions.

3.3.2. Field experiments

A complementary approach to measuring and disentangling the nature of discrimination is to use field experiments. Although a very recent study thoroughly catalogues a variety of field experiments that test for discrimination in the marketplace (Riach and Rich, 2002), a brief summary of the empirical results is worthwhile to provide a useful benchmark. Labor market field studies present perhaps the broadest line of work in the area of discrimination. The work in this area can be parsed into two distinct categories: personal approaches and written applications.

Personal approaches include studies that have individuals either attend job interviews or apply for employment over the telephone. In these studies, the researcher matches two testers who are identical along all relevant employment characteristics except the comparative static of interest (e.g., race, gender, age). Then, after appropriate training, the testers approach potential employers who have advertised a job opening. Researchers "train" the subjects simultaneously to ensure that their behavior and approach to the job interview are similar.

Under the written application approach, which can be traced to Jowell and Prescott-Clarke (1970), carefully prepared written job applications are sent to employers who have advertised vacancies. The usual approach is to choose advertisements in daily newspapers within some geographic area to test for discrimination. Akin to the personal approaches, great care is typically taken to ensure that the applications are similar across several dimensions except the variable of interest.

It is fair to say that this set of studies, including both personal and written approaches, has provided evidence that discrimination against minorities across gender, race, and

age dimensions exists in the labor market. But due to productivity unobservables, the nature or cause of discrimination is not discernible. This point is made quite starkly in Heckman and Siegelman (1993, p. 224), who note that "audit studies are crucially dependent on an unstated hypothesis: that the distributions of unobserved (by the testers) productivity characteristics of majority and minority worker are identical." They further note (p. 255): "From audit studies, one cannot distinguish variability in unobservables from discrimination." Accordingly, while these studies provide invaluable insights into documenting that discrimination exists, care should be taken in making inference about the type of discrimination observed.

Much like the labor market regression studies discussed above, the literature examining discrimination in product markets has yielded important insights. Again, rather than provide a broad summary of the received results, we point the reader to Yinger (1998) and Riach and Rich (2002), who provide nice reviews of the product market studies.[38] We would be remiss, however, not to at least briefly discuss the flavor of this literature.

One often cited, recent study is the careful work due to Bertrand and Mullainathan (2004), who utilize a natural field experiment to determine whether or not blacks are discriminated against by employers. By sending resumes with randomly assigned white- or black-sounding names to want-ads advertised in Boston and Chicago newspapers, Bertrand and Mullainathan find that white names receive 50% more callbacks for an interview than black names. This racial gap is uniform across occupation, industry, and employer size. Additionally, whites receive greater benefits to a higher-quality resume than blacks. Although Bertrand and Mullainathan are unable to test the type of discrimination, whether taste-based or statistical, as it is uncertain what information the employer is utilizing from the resumes, the authors use the results to suggest an alternate theory be considered, such as one based on lexicographic searches.

To choose names that are distinctly white-sounding or black-sounding, Bertrand and Mullainathan use name frequency data calculated from birth certificates of all babies born in Massachusetts between 1974 and 1979. Distinctiveness of a name is calculated as having a sufficiently high ratio of frequency in one racial group to that of the other racial group. The 9 most distinct male and 9 most distinct female names for each racial group, along with corresponding white- or black-sounding last names, are used. To verify this method of distinction, a brief survey was conducted in Chicago asking respondents to identify each name as "White", "African-American", "Other", or "Cannot Tell." Names that were not readily identified as white or black were discarded.

The authors sampled resumes posted more than six months prior to the start of the experiment on two job search websites to use as a basis for experimental resumes. The

[38] The interested reader should also see the recent special Symposium issue on Discrimination in Product, Credit, and Labor Markets that appeared in the Journal of Economic Perspectives Spring (1998).

resumes sampled were restricted to people seeking employment in sales, administrative support, clerical services, and customer service in Boston and Chicago, and were purged of the original owner's name and address. To minimize similarities to actual job seekers, Chicago resumes are used in Boston and Boston resumes are used in Chicago (after names of previous employers and schools are changed appropriately). The quality of the resumes were sorted into two groups (high and low), with high-quality resumes having some combination of more labor market experience; fewer gaps in employment history; being more likely to have an e-mail address, certification degree, or foreign language skills; or been awarded honors of some kind. Education is not varied between high- and low-quality resumes to ensure each resume qualifies for the position offered, and approximately 70% of all resumes included a college degree of some kind.

Fictitious addresses were created and randomly assigned to the resumes based on real streets in Boston and Chicago. The authors selected up to three addresses in each 5-digit zip code in both cities using the White Pages. Virtual phone lines with voice mailboxes were assigned to applicants in each race/sex/city/resume quality cell to track callbacks. The outgoing message for each line was recorded by someone of the appropriate race and gender, and did not include a name. Additionally, four e-mail addresses were created for each city, and were applied almost exclusively to the high-quality resumes.

The field experiment was carried out between July 2001 and January 2002 in Boston and between July 2001 and May 2002 in Chicago. In most cases, two each of the high- and low-quality resumes were sent to each sales, administrative support, and clerical and customer services help-wanted ad in the Sunday editions of The Boston Globe and The Chicago Tribune (excluding ads asking applicants to call or appear in person to apply). The authors logged the name and contact information for each qualifying employer, along with information on the position advertised and any specific requirements applicants must have. Also recorded was whether or not the ad explicitly stated that the employer is an "Equal Opportunity Employer."

For each ad, one high-quality resume and one low-quality resume were randomly assigned a black-sounding name (with the remaining two resumes receiving white-sounding names). Male and female names were randomly assigned for sales jobs, while primarily female names were used for administrative and clerical jobs to increase the rates of callbacks. Addresses were also randomly assigned, and appropriate phone numbers were added before formatting the resumes (with randomly chosen fonts, layout, and cover letters) and faxing or mailing them to the employer. A total of 4870 resumes were sent to over 1300 employment ads. Of these, 2446 were of high-quality while 2424 were of low-quality.

Results are measured by whether a given resume elicits a callback or an e-mail back for an interview. Resumes with white-sounding names have a 9.65% chance of receiving a callback compared to 6.45% for black-sounding names, a 3.2 percentage point difference. This difference can only be attributed to name manipulation. According to

these results, whites are 49% (50%) more likely to receive a callback for an interview in Chicago (Boston). This gap exists for both males and females, with a larger, though statistically insignificant, racial gap among males in sales occupations. An additional year of workforce experience increases the likelihood of a callback by approximately 0.4 percentage point, thus the return to a white name is equivalent to 8 additional years of experience. High-quality resumes receive significantly more callbacks for whites (11% compared to 8.5%, $p = 0.0557$), while blacks only see a 0.51% increase (from 6.2% to 6.7%). Whites are favored (defined as more whites than blacks being called back for a specific job opening) by 8.4% of employers, where blacks are favored by only 3.5% of employers, a very statistically significant difference ($p = 0.0000$). The remaining 88% of employers treat both races equally, with 83% of employers contacting none of the applicants.

A probit regression of the callback dummy on resume characteristics (college degree, years experience, volunteer experience, military experience, e-mail address, employment holes, work in school, honors, computer skills, special skills, fraction of high school dropouts in the neighborhood, fraction of neighborhood attending college or more, fraction of neighborhood that is white, fraction of neighborhood that is black, and log median per capital income) is created from a random subsample of one-third of the resumes. The remaining resumes are ranked using the estimated coefficients by predicted callback. Under this classification, blacks do significantly benefit from high-quality resumes, but they benefit less than whites (callback rates for high versus low are 1.6 for blacks and 1.89 for whites). The presence of an e-mail address, honors, or special skills have a positive significant effect on the likelihood of a callback. Interestingly, computer skills negatively predict callback and employment holes positively predict callback. Additionally, there is little systematic relationship between job requirements and the racial gap in callback.

Applicants living in whiter, more educated, or higher-income neighborhoods have a higher probability of receiving a callback, and there is no evidence that blacks benefit any more than whites from living in a whiter, more educated zip code. There is, however, a marginally significant positive effect of employer location on black callbacks.

Of all employers, 29% state that they are "Equal Opportunity Employers" and 11% are federal contractors, however these two groups are associated with a larger racial gap in callback. The positive white/black gap in callbacks was found in all occupation and industry categories except for transportation and communication. No systematic relationship between occupation earnings and the racial gap in callback was found.

Bertrand and Mullainathan did not design their study specifically test the two theories of discrimination, statistical and taste-based, and do not believe that either of the two can fully explain their findings. While both models explain the average racial gap, their results do not support animus. There is no evidence of a larger racial gap among jobs that explicitly require communication skills and jobs for which customer or

co-worker contacts are more likely to be higher, which would be expected by theory. Further, as blacks' credentials increase the cost of discrimination should increase, but this doesn't explain why blacks get relatively lower returns to a higher-quality resume. This, combined with the uniformity of the racial gap across occupations, casts doubt on statistical discrimination theories as well.

The authors suggest that other models may do a better job than statistical or taste models at explaining these particular findings. For example, a lexicographic search by employers may result in resumes being rejected as soon as they see a black name, thus experience and skills are not rewarded because they are never seen. This theory may explain the uniformity of the race gap if this screening process is similar across jobs. The results could also follow from employers having coarser stereotypes for blacks. In any case, Bertrand and Mullainathan acknowledge the need for a theory beyond statistical discrimination and taste to explain their findings in full.

Another nice example of a natural field experiment is due to Riach and Rich (2006), who extend the literature by using carefully matched written applications made to advertised job vacancies in England to test for sexual discrimination in hiring. They find statistically significant discrimination against men in the "female occupation" and against women in the "male occupation." This is important evidence to begin to uncover the underlying causes for labor market discrimination. This study is also careful to point out that it is difficult to parse the underlying motivation for why such discrimination exists. Even without such evidence, however, the paper is powerful in that it provides a glimpse of an important phenomenon in a significant market, and provocatively leads to questions that need to be addressed before strong policy advice can be given.

There are a number of other studies that examine discrimination and differential earnings in labor markets based on sexual orientation (Arabsheibani et al., 2005; Weichselbaumer, 2003; Berg and Lien, 2002), but like these two natural field experiments, they also have difficulties parsing the type of discrimination observed.

One might then ask, if field experiments have similar difficulties as regression based methods in parsing the nature of discrimination, why bother with this approach. Our answer is that field experiments in labor economics have the potential to parse both the nature and extent of discrimination observed in markets.

As a starting point, consider List (2004b), who made use of several settings in a naturally-occurring marketplace (the sports card market) to show that a series of field experiments can parse the two forms of discrimination. More specifically, after first demonstrating that dealers treat "majority" (white men) and "minority" (older white men, nonwhite men and white women) buyers and sellers in the marketplace differently, List provides evidence suggesting that sportscard dealers knowingly statistically discriminate. By executing a variety of field experiments, the evidence provided parses statistical discrimination from taste based discrimination and an agent's ability to bargain when interacting with a dealer. The experiments conducted by List demonstrate a

framework for potentially parsing the two forms of discrimination which could be utilized and moved forward to inform discrimination discussions in other markets.

The first experiment discussed in List (2004b) is similar to an audit study in that dealers are approached by buyers from both majority and minority groups with an offer of buying or selling a sportscard (unlike most audit studies, the subjects do not know that they are part of a study on discrimination, just that it's an economic study). The results from this first experiment are highly suggestive that dealers base offers on group membership: buyers in the minority groups of white women and older white men received initial offers that were 10-13% greater than white male buyers when buying cards and minority groups received 30% lower initial offers when selling their cards.

Further, this initial framed field experiment shows that the gap between minority and majority subjects' offers remain from the initial to final offers for inexperienced subjects but to a large part converges for subjects with experience. But, this convergence comes at a cost of time: subjects in the minority group having to invest a significantly larger portion of time to achieve better final offers.

This result provides support for non-taste based discrimination due to the convergence of the gap in offers through bargaining, a result that would not hold under a theory of taste based discrimination where the dealer would simply hold to one price. Finally, by surveying dealers in addition to subjects that were buying and selling, List controls for dealer experience in the marketplace and finds a positive relationship between dealer experience and discrimination as measured by the difference between a dealer's average majority and average minority offers. Suggesting that statistical discrimination may be evident unless one believes that taste based discrimination increases with experience as a dealer at sportscard shows.

Although this initial experiment can measure discrimination, more treatments are necessary to parse statistical discrimination from alternative explanations. In total, List runs four experiments in addition to the framed field experiment described in the previous paragraph: (i) a dictator game artefactual field experiment with dealers as the dictator and four descriptions as the receiver: white men, non-white men, white women and white mature men; (ii) two framed field experiment treatments that are bilateral exchange markets with dealers selling to agents with randomly drawn reservations values, where in one market dealers know that the reservation value is random and in a second it is ambiguous; (iii) a Vickrey second price auction that is a natural field experiment; and, (iv) a framed field experimental game designed to determine dealers' perceptions of the reservation value distributions of sportscard market participants. Each additional experiment helps parse the two forms of discrimination and the bargaining ability of the subjects and the results of all the experiments are necessary for List to suggest that dealers knowingly statistically discriminate.

First, the relatively uniform offers made to receivers across majority and minority groups in the dictator game suggests that dealers do not display taste based discrimination,

at least in artefactual field experiments. Second, through the bilateral exchange markets, three results are found which each point towards statistical discrimination by testing hypotheses drawn directly from the theories of taste based and statistical discrimination. First, experienced dealers are found to lose less surplus than inexperienced dealers. Second, minority and majority buyers perform similarly with the randomly set reservation prices but not in the treatment where dealers think that reservation values are "homegrown values." Finally, experienced dealers perform worse when it is ambiguous whether the reservation value is drawn randomly—suggesting that they are utilizing inferences which are not performing well (i.e. their statistical discrimination rubric fails due to the randomly set reservation value). These two additional experiments point toward statistical discrimination.

Yet, it is only through the final two experiments that sufficient evidence is provided for statistical discrimination through a discovery of a variation in reservation value distributions of sportscard market participants and dealer knowledge of the variation. The results from the Vickrey second price auction are used for two purposes: (i) to determine whether the reservation value distributions of the majority and minatory are indeed different and (ii) to provide distributions to determine the dealers' abilities to accurately assign distributions in the final experimental game. The results from the Vickrey auction do show that the reservation values for the minority group have a larger variance than the reservation values for the majority group, suggesting that statistical discrimination could be utilized for profit maximization—see the above model. Further, when different reservation value distributions are shown to dealers in the final experiment, a majority of all dealers are able to determine which distributions are from which groups and experienced dealers are able to correctly assign distributions more often than inexperienced dealers.

Although List focuses on a market that every consumer does not necessarily approach, the framework of multiple field experiments to move towards identifying the form of statistical discrimination is one that should be considered for use elsewhere. Most importantly, this study highlights that a series of field experiments can be used to uncover the causes and underlying conditions necessary to produce data patterns observed in the lab or in uncontrolled field data.

This study shows highlights that a deeper economic understanding is possible by taking advantage of the myriad settings in which economic phenomena present themselves. In this case, field experimentation in a small-scale field setting is quite useful in developing a first understanding when observational data is limited or experimentation in more "important" markets is not possible. Yet, it is important to extend this sort of analysis to more distant domains.

This is exactly what is offered in Gneezy et al. (2010), who explore the incidence of discrimination against the disabled by examining actual behavior in a well-functioning marketplace—the automobile repair market. This study uses a traditional audit study, but

combines it with a specific field experimental treatment to allow the authors to parse the type of discrimination observed.

The audit portion of the study was standard: the assignment given to subjects is clear: approach body shop j to receive a price quote to fix automobile i. The authors included subjects from two distinct groups—disabled white males age 29-45 and non-disabled white males age 29-45—who each visited six body shops. The disabled subjects in this experiment were all confined to a wheelchair and drove a specialized vehicle. All of the automobiles, which were personally owned by our disabled subjects, had visible body problems. Importantly, both testers in any given pair approached body shops with the identical car.

The authors find that overall, the disabled received considerably higher average price quotes, $1425, than the non-disabled, $1212. Inference as to why this disparate treatment exists, of course, is an open issue. Several clues provide potential factors at work: (i) access—many body shops are not easily approachable via wheelchairs; this considerably restricts the set of price offers the disabled can receive; and (ii) time—while the non-disabled can easily park and proceed to the front desk, the process is much more complex for the disabled. First, he must find a suitable parking place: it is very uncommon to have designated places for the disabled in body shops. As a result, the disabled must have special parking which permits the use of a wheelchair. Moreover, it must be a space that will be unoccupied when he returns to pick up the repaired vehicle. After finding an appropriate parking space, the disabled must commit much more effort and time to approach the service desk. An additional related problem which makes the expected search cost higher for the disabled person is that in some cases it is necessary to leave the car for the day in order to obtain a price quote. Using a taxi is much more complex for the disabled than for the non-disabled.

To investigate the search cost explanation further, the authors obtained data on search effort at the tester level and perceived search effort at the body shop level to examine if realizations of these variables are consistent with the pattern of discrimination observed. From this survey, the authors find that the non-disabled typically consult far fewer body shops: on average, the non-disabled visit 3.5 different mechanics whereas the disabled visit only 1.67 mechanics, a difference that is statistically significant. Concerning the supply side, the authors asked body shops questions revolving around body shop perceptions of the degree of search among the disabled and non-disabled. The results are consonant with the consumer-side statements observed above: the disabled are believed to approach 1.85 body shops for price quotes while the non-disabled are expected to approach 2.85, a difference of more than 50 percent and one that is significant. This evidence is consistent with statistical discrimination based on mechanics' beliefs about relative search costs and how they map into reservation value distributions. Yet, the survey evidence alone is only suggestive and further investigation is necessary to pinpoint the underlying mechanism at work.

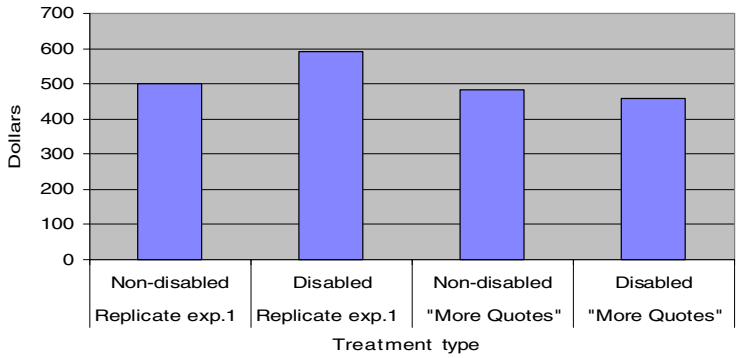


Figure 8 ***Complementary experiment I summary.***

For these purposes, the authors provide a sharper focus on the underlying reason for discrimination by running a complementary field experiment. In this experiment, the authors not only replicated the initial results, with new testers and new vehicles in need of repair, but in another treatment had these exact same agents approach body shops explicitly noting that [they're] "getting a few price quotes" when inquiring about the damage repair estimate. If differential search costs cause discrimination, then the authors should observe the offer discrepancies disappearing in this treatment.

This is exactly the result they observe. Although in the replication treatment the disabled continued to receive higher asks, when both agent types noted that they were "getting a few price quotes," the disabled agents were able to secure offers that were not statistically distinguishable from the offers received by the non-disabled. We provide support of this insight in Fig. 8, which highlights the discrepancies observed when search is believed to be heterogeneous across the disabled and non-disabled. In this case, the first two bars show that the differences are nearly 20%. Yet when both agents clearly signal that this particular mechanic visit is just one part of their entire search process, these disparities are attenuated and indeed change signs.

While these two examples are not directly related to labor market outcomes, they display the power of the field experimental method to test important theories within labor economics, and especially the theories of discrimination discussed earlier. In this regard, we believe that similar treatments can be carried out in labor markets to explore wage differences, job offer differences, and other labor market outcomes.

4. FIRMS

In the third stage of the life cycle, most individuals work within firms or some other hierarchical organization. In this section we describe how field experiments have contributed to knowledge on how workers and employees behave in such settings. Following the research areas described in Tables 2 and 3, we frame our discussion on the

following research themes: (i) the effects on monetary incentives on worker behavior; (ii) the interplay between monetary and non-monetary incentives; (iii) aspects of the employment relationship, such as gift-exchange between employers and employees, and the link between employer monitoring and employee shirking behavior.[39, 40]

4.1. Monetary incentives

A core principle in economics is that incentives matter. The role of monetary incentives within firms and organizations has been long studied in sociology and management literatures. With the application of contract theory to behavior within firms (Hart and Holmstrom, 1987) and the development of personnel economics (Lazear, 1995), such questions are now integrated within mainstream labor economics. For economists, the basic questions have been: (i) how do workers respond to a given set of incentives?; (ii) what are the optimal set of incentives an employer should provide?[41]

An earlier generation of empirical studies exploited firm's personnel data to measure the productivity effects of compensation schemes on individual workers. An econometric challenge facing these studies is that observed incentive contracts might well be endogenous to firm's performance (Prendergast, 1999; Chiappori and Salanie, 2003). In other words, identifying causal effects of incentives on behavior is confounded by the presence of unobservables, such as managerial practices, that determine both which compensation schemes are chosen, and worker productivity. In earlier research this concern has been addressed in between firm studies using instrumental variables approaches (Groves et al., 1994). However, this concern applies even if such effects are identified from a within worker or within firm comparison as incentives change over time (Jones and Kato, 1995; Ichniowski et al., 1997; Paarsch and Shearer, 1999, 2000; Lazear, 2000). A related concern is that such changes in incentives might be reflective of a wider package of changes in management practices. Hence, akin to

[39] The field experiment approach shares many of the characteristics of the insider econometrics approach to understand the causes and consequences of behavior within a firm (Ichniowski and Shaw, forthcoming). However a key distinction is that field experiments explicitly rely on exogenous variation created with the specific influence of researchers in order to identify causal effects. Clearly, not every intervention that a researcher could design and implement is socially useful—there is little value added in implementing practices that firms are never otherwise observed engaging in. However, this does not preclude the fact that carefully designed interventions can help researchers to uncover causal relations and the mechanisms behind them.

[40] Our discussion focuses predominantly on natural field experiments within firms. There also exists a separate branch of artefactual field experiments where subject pools are drawn from manufacturing workers (Barr and Serneels, 2009), fishermen (Carpenter and Seki, 2010) and employees in large firms (Charness and Villeval, 2009).

[41] Many of the wider literature related to the research questions we touch upon, such as incentive pay and teams, are discussed in greater detail in the Chapter on Human Resource Management by Bloom and Van Reenen (2011), also in this Handbook. They summarize the evidence from across countries showing the increasing use of performance pay over time. In the Chapter on Personnel Economics in this Volume by Oyer and Schaefer (2011), further issues related to incentive pay and firm hires is discussed at greater length.

social experiments, what is actually being evaluated is potentially the sum total of many concomitant changes in the firm's organization rather than an isolated change in worker incentives all else equal. This is of particular concern given the view that there exist complementarities between organizational practices so that firms are better off choosing a package of practices rather than in isolation (Milgrom and Roberts, 1990; Ichniowski et al., 1997). With such multiple underlying changes, mapping the evidence, however cleanly identified is the change in behavior, to any underlying theory is less clear cut.[42]

Field experiments introduce exogenously timed variation in incentive structures that are orthogonal to other management practices. This opens up the possibility to identify the causal impact of monetary incentives on the behavior of individual workers, and on firm performance as a whole. Combining personnel files from human resource departments within the firm, with primary data collection that is inherent in field experimentation, allows researchers to examine the effect of monetary incentives on a range of margins of worker behavior, capturing both the intended and unintended consequences of incentive provision.

There are good theoretical reasons for collecting such extensive information on worker behaviors when evaluating the response to incentives. For example, multi-tasking theory suggests that when monetary incentives are provided based on a subset of tasks that the firm can directly measure performance in, workers may reallocate their effort away from other tasks they are engaged in, their employer is affected by, but their compensation is not based on (Holmstrom and Milgrom, 1991). Similarly, if the provision of incentives alters the distribution of pay across workers in the same tier of the firm hierarchy, this might alter worker's behavior towards co-workers, say through cooperation or sabotage (Lazear, 1989). Finally, there might be ways in which workers can game against any incentive scheme. All such unintended consequences of monetary incentives need to be accounted for to both accurately understand how workers respond to incentives and to begin to think through the optimal incentive design.[43]

Employers might not collect such information *ex ante*. Hence the need to engage in primary data collection efforts to complement the rich information available in firm's personnel files. Field experiments—that involve close cooperation between researchers and firm management—are well placed to advance in this direction. Ultimately, as witnessed in some of the field experiments described below, this allows a closer mapping between the evidence and underlying theory, and to draw implications for optimal incentive provision.

[42] Due to these empirical challenges, it is not surprising that much of the early evidence testing theories in personnel economics originated from laboratory environments. For example, Bull et al. (1987) provide evidence from the lab on the predictions of rank order tournament theory; Fehr and Fischbacher (2002) review the experimental evidence on social preferences in workplace environments. The wider availability of personnel data and ever closer links being forged between researchers and firms has allowed the literature in field experiments within firms to flourish.

[43] Charness and Kuhn (2011, 2010) review the extensive evidence from laboratory settings on sabotage.

4.1.1. Theoretical framework

To understand some of the theoretical questions and empirical challenges faced in this literature, it is instructive to first reconsider Lazear's (2000) original analysis of the Safelite Glass Corporation, a large auto-glass firm in which the primary task of worker's at the bottom-tier of the firm's hierarchy is to install automobile windshields. Lazear used non-experimental methods to estimate the productivity effects of the firm moving from a compensation scheme in which workers were paid an hourly wage scheme, to one in which they were paid a piece rate for each windshield installed, with a minimum guarantee. This pioneering work brings to the fore many of the issues that have influenced all the subsequent literature, and allows us to highlight the specific issues that field experiments help address.

The model is as follows. Worker's utility depends on income Y and effort e, $U(Y, e)$ with $U_1 > 0, U_2 < 0$.[44] Worker's output q depends on effort and her ability, θ, so $q = q(e, \theta)$ with output assumed to be observable and $q_1, q_2 > 0$. For any given output q_0, there is a unique effort level that achieves this, denoted $e_0(\theta)$. It is then straightforward to see that $\frac{\partial e}{\partial \theta} = -\frac{q_2}{q_1} < 0$ so that higher ability workers need exert less effort to achieve a given output. If workers choose not to work at any firm, their outside option from leisure is denoted $U(0, 0)$. Hence the lowest ability worker that would accept employment at a firm with a required output level and wage W, is denoted θ_0 and is such that,

$$U(W, e_0(\theta_0)) = U(0, 0). \tag{27}$$

All workers of higher ability earn rents from employment over leisure. Similarly, suppose a worker of a given ability could take up employment at another firm offering a wage-minimum effort pair $(\widehat{W}, \widehat{e})$. Hence with inter-firm competition there might exist an upper cutoff in ability, θ_h, such that,

$$U(W, e_0(\theta_h)) = U(\widehat{W}(\theta_h), \widehat{e}(\theta_h)), \tag{28}$$

where workers of ability higher than θ_h prefer to take the alternative employment contract.

[44] There is a long-standing idea in psychology that rewards may hinder performance (Kruglanski, 1978). There is some evidence on this from laboratory settings where offering small amounts of monetary compensation is found to decrease effort relative to paying nothing (Gneezy and Rustichini, 2000), and where explicit incentives sometimes result in worse compliance than incomplete labor contracts (Fehr and Falk, 1999; Fehr and Schmidt, 2000). This might either be because small monetary incentives crowd out intrinsic motivation, an idea formalized by Benabou and Tirole (2000), or because the individual is reluctant to signal his willingness to accept low wages. We do not know of field evidence that examines such non-monotonic effects of monetary incentives on effort.

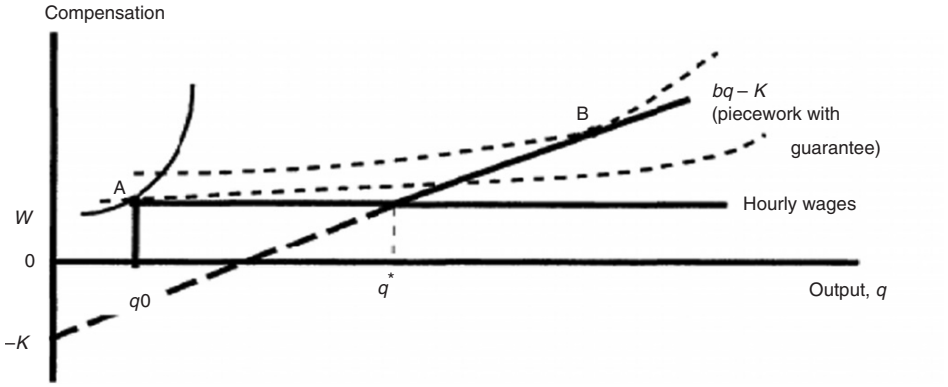


Figure 9 ***Compensation before and after at Safelite.***

This framework makes clear that incentive structures will affect two types of behavior, an idea developed in more detail in Lazear (2005). First, there will be change in effort e exerted by individual workers in response to monetary incentives. This is referred to as the "incentive effect". Second, the compensation scheme will induce a differential composition of workers within the firm over time. Some workers will prefer to join this firm from other employers. These changes in workforce composition can be thought of as the "selection effect" of monetary incentives.

The incentive effect can be easily understood graphically. Figure 9 shows the relationship between output, q, and compensation for the two schemes relevant for Lazear's study: (i) a fixed hourly wage subject to a minimum output requirement q_0, resulting in total compensation W; (ii) a linear piece rate scheme $bq - K$ with a minimum guarantee of W. As Fig. 9 shows, for output levels between q_0 and q^* the worker receives W under both compensation schemes, and for output higher than q^* earns strictly more under the piece rate scheme.[45]

On the incentive effect, the model makes clear that moving from the fixed hourly wage scheme to the piece rate scheme does not cause the output of any individual to fall, and causes average output to rise. Low ability workers, indicated with a solid indifference curve in Fig. 9, remain indifferent between the two schemes and would produce output q_0 at point A under both. Higher ability workers, indicated with a dashed indifference curve in Fig. 9, would prefer to increase their effort and move to point B. This is because the piece rate scheme allows higher ability workers to raise their utility through increased compensation that more than offsets any increase in their effort. As a result, the dispersion

[45] Firms typically provide workers some insurance by allowing their output to occasionally fall below the required minimum q^*, but a worker that consistently fails to meet this performance threshold is likely to be fired or assigned to another task.

in worker effort and output rises as long as there is at least one worker that chooses to produce more than q^*.[46]

On the selection effect, under plausible conditions, the average ability of workers rises with the move to piece rates. This is because low ability individuals remain indifferent between working for this firm under either incentive scheme. If they were willing to work for the firm under the fixed wage scheme, they should remain willing to do so under piece rates all else equal. On the other hand, high ability workers might be attracted to this form from other firms that for example, have higher minimum output standards or pay piece rates but at a lower rate b. In short, theory predicts that there should be no change in the number of low ability workers who are willing to work at the firm, but that piece rates attract high ability workers so the right tail of the ability distribution in the firm should thicken.

As described in more detail below, existing field experiments have focused on identifying the incentive effects, and the research designs used have been less amenable to pin down these types of selection effect. Yet it is important to emphasize the need for future research to provide credible research designs to uncover both effects.[47]

In Lazear's study, he documents the total effect of the change in monetary incentives to workers was around a 44% increase in worker productivity, defined to be the number of windshields installed by the worker per eight hour day. Around half the increase was due to incentive effects, namely a change in effort of the same worker as he moved from a fixed hourly wage to a piece rate scheme. However, the other half was entirely due to the selection effect, namely productivity changes due to endogenous changes in the composition of workers in response to change in monetary incentives. A legacy of Lazear's study is to show that both motives likely underlie why firms choose to alter their output based incentive structures in the first place. A carefully crafted field experiment that begins to measure whether and how the compensation policies of a given firm have such spillover effects on other firms that compete for similar workers, would open up a rich research agenda tieing together the study of within-firm compensation policies on equilibrium wage-setting behavior in labor markets.[48]

[46] Whether workers exert more or less effort in response to a higher piece rate b of course depends on the balance of income and substitution effects. Evidence from the lab and field in Gneezy and Rustichini (2000) suggested that the relationship between piece rates and effort was U-shaped with low piece rates eliciting less effort than a zero piece rate. One explanation would be that small levels of financial compensation crowd out workers' intrinsic motivation to exert effort.

[47] Laboratory experiments have begun to explore in more detail the selection effects of incentives (Dohmen and Falk, 2006; Cadsby et al., 2007; Vandegrift et al., 2007; Niederle and Vesterlund, 2007; Eriksson et al., 2008a,b). These studies are described in more detail in Charness and Kuhn (2011, 2010).

[48] This links together with recent development in structural estimation of search and matching models in labor markets. For example, Cahuc et al. (2006) develop and estimate an equilibrium model with strategic wage bargaining and on-the-job search. An important innovation on their paper is that when an employed worker receives an outside job offer, a three-player bargaining process is started between the worker, her/his initial employer and the employer which made the outside offer. They use the model to examine wage determination in France using matched employer-employee

Linking to the design of field experiments

The model provides a series of implications that have impinged on the first generation of field experiments over the last decade. First, given worker heterogeneity, changes in compensation scheme will nearly always affect average effort and output, as well as the dispersion of effort and output. Given the linkage between performance and pay, this inevitably results in changes in the distribution of earnings across workers at the same tier of the firm hierarchy. Hence linking pay to performance might have unintended negative consequences on worker and firm performance as a result of such increased earnings inequality. This might manifest itself in the form of workers reducing cooperation with co-workers (Baron and Pfeffer, 1994; Bewley, 1999; Lazear, 1989), workers sabotaging the performance of others, or workers being directly worse off in utility terms, all else equal, as a result of them being structurally averse to pay inequality (Fehr and Schmidt, 1999; Charness and Rabin, 2002). Field experiments are particularly adept at detecting and quantifying such unintended consequences because researchers are engaged in primary data collection, and the firm would have no incentive to collect such information *ex ante* as part of its personnel files, especially if pay for performance type compensation schemes have not been previously implemented. Some of the field experiments described below have collected qualitative evidence from workers to explore these channels, in additional to using personnel files to measure the direct productivity effects of incentives.

Second, given the selection effect of monetary incentives, there is inevitably a change in workers' peer group over time. The composition of peers, or their social ties with each other, play no role in the standard neoclassical model in which worker preferences are only defined over their own income and effort. There are good reasons to probe this assumption in the field. First, if peer effects determine workplace behavior because they alter the marginal return to worker's effort, then understanding how workers respond to changes in monetary incentives requires an understanding of the mechanisms underlying such peer effects. Second, extending the neoclassical model to take into account such peer effects or social concerns, as has been done in Kandel and Lazear (1992), Lazear (1989), Rotemberg (1994) and Fershtman et al. (2003), has implications for many aspects of firm behavior including the optimal design of incentives. Such concerns are a recurring theme in the series of natural field experiments conducted by Bandiera et al. that are described below.

Third, the model highlights that under fixed hourly wage schemes, there should not be much heterogeneity in workers output or effort, despite workers being heterogeneous in ability. This does not fit the evidence very well. For example, in Lazear's study there was considerable dispersion in worker productivity even under the fixed hourly wage scheme. To better explain behavior in fixed wage settings, theory suggests workers might

data from 1993 to 2000. They find that inter-firm competition is quantitatively important for wage determination, and raising wages above reservation levels.

respond to other forms of non-monetary or implicit incentives, such as gift-exchange motives where workers exert more effort in response to employers paying higher than market clearing wages, the ability to shirk (Shapiro and Stiglitz, 1984; Macleod and Malcomson, 1989), or promotion prospects and career concerns (Dewatripont et al., 1999). Such mechanisms might also predict how workers sort across firms (Stiglitz, 1975). Some of these aspects are highlighted by the field experiments discussed below on the employment relationship.

Fourth, the model implies that low ability workers should not leave the firm with the move to piece rates. If they were willing to work for the firm under the fixed wage scheme, they should remain willing to do so under piece rates all else equal. On the other hand, given output differences among workers under price rates, management can more easily identify low ability workers when pay is tied to their performance. Over a longer time period, this might lead to them being fired. This type of selection effect caused by employer learning the true ability of workers, has not been studied by field experiments.[49]

Finally, given worker heterogeneity, the first best for the firm would be to set a worker specific piece rate, b_i. This would be chosen to equate the marginal benefit of effort to its marginal cost. However, we generally observe firms being constrained to offer bottom-tier workers the *same* compensation scheme. This may be because of legal, technological or informational constraints (Lazear, 1989; Bewley, 1999; Encinosa et al., 1997; Fehr et al., 2004). To overcome this, one hypothesis is that firms can get closer to this first best by linking managers' pay to the firm's performance. Managers then have greater incentives to target their effort to specific workers, and from the worker's point of view it is then as if they face an individual specific incentive scheme. This idea is developed in the natural field experiment by Bandiera et al. (2007) described below, which then links managerial incentives to pay inequality among workers.

4.1.2. Evidence from the field

In the decade following the wave of studies using personnel data and insider econometrics to understand responses to monetary incentives (Ichniowski and Shaw, forthcoming), field experiments have begun to exploiting exogenous and randomly timed variation in compensation pay to both reinforce these existing results using non-experimental methods, as well as providing new insights.

Among the first of these studies was Shearer's (2004) natural field experiment, designed to estimate the productivity gains moving from a piece rate to a fixed wage

[49] A related concern has been on the existence of rachet effects in response to pay for performance (Gibbons, 1987), whereby workers deliberately underperform to keep the piece rate high. Such ratchet concerns have been documented in firms where productivity shocks are uncommon such as shoe making (Freeman and Kleiner, 2005) and bricklaying (Roy, 1952). Cooper et al. (1999) present evidence from an artefactual field experiment on Chinese students and managers on such ratchet concerns, that might be of particular concern in planned economies.

scheme for tree planters in British Columbia, Canada. In contrast to Lazear (2000), this setting is one in which tree planters are usually paid a piece rate with no guaranteed minimum, and fixed wages are rarely used *ex ante*. Workers were randomly assigned at the start of a work day, to plant under one of the incentive schemes. Hence a within worker comparison can be exploited to estimate the incentive effects. To reduce the likelihood of this comparison confounding other effects unrelated to the compensation schemes in place, work took place on fields of similar conditions, and there was a constant length of the work day. The incentive effect estimate is then based on a total of 120 observations on daily worker productivity, 60 under each scheme. The relatively small sample size—nine male workers were randomly selected from the firm—reflects the difficulty researchers initially faced in real world settings in convincing firms to randomly assign individuals to alternative compensation schemes. As more field experiments are conducted, some of these constraints are being eased. For example, some of the field experiments described below are based on data on hundreds of workers.[50]

Shearer's field experiment reveals the incentive effect of having a piece rate rather than a fixed wage compensation schemes to be a 20% productivity increase. The magnitude of this is comparable to Lazear's (2000) findings—moving from a fixed wage scheme—based on non-experimental data, although this is wholly by chance There is no reason *a priori* to expect the behavioral response of workers to these two incentive schemes to be of the same magnitude given the very different types of worker involved, nature of the production function, and that the piece rate b was not the same across settings. In line with theory, Shearer finds the standard deviation of output across workers was higher under piece rates. Overall, it was found that unit costs under piece rates were around 13% lower than under fixed wages.

To shed more light on how workers would have responded in a slightly different environment and to alternative compensation schemes, Shearer then develops and estimates a structural model. In terms of altering the economic environment, the structural model is used to shed light on what would have been the productivity gains if management was imperfectly informed about planting conditions. This yields comparable estimates of the productivity gains of moving to piece rates. To shed light on alternative compensation schemes, Shearer explores how workers would have responded to an efficiency wage scheme. The efficiency wage that would induce effort levels equal to those observed under fixed wages in the field experiment, is calculated and its implied unit costs are compared with those achieved under piece rates. This exercise suggests that fixed wages would lead to a 2.7% increase in unit costs relative to piece rates.

Shearer (2004) uses a close to best practice methodology in combining estimates from a field experiment with structural modelling within the same setting. This combination,

[50] The small sample size used in Shearer (2004) also reflects the nature of tree planting firms. They typically employ less than 100 planters.

following from discussions in Heckman and Smith (1995) and Keane and Wolpin (1997), first identifies the existence and magnitude of important causal effects using reduced form evidence. To then move away from such black-box findings, the researcher then uses structural modelling to posit an underlying behavioral mechanism behind the effects, assess the sensitivity of the estimates to slight alterations in the economic environment, and to make headway in understanding the optimal compensation structure. Of course, the validity of the structural model can itself be tested by exploring whether it predicts the responses observed to the exogenous variation engineered by the field experiment.

Another example of how field experiments can and should learn from other methodologies is at the heart of the natural field experiment of Hossain and List (2009). They use theoretical insights on framing effects from behavioral economics that have previously found empirical support in laboratory experiments (Kahneman and Tversky, 1979; Thaler, 1980; Samuelson and Zeckhauser, 1988; Ellingsen and Johannesson, 2008), to see if, in the field, framing manipulations affect worker responses to bonus incentives. Their setting is a high tech Chinese firm producing consumer electronics, where workers are organized into both individual and team production. They find that bonuses framed as both "losses" and "gains" increase productivity, for both individuals and teams. Teams respond more to bonuses posed as losses than as comparable bonuses posed as gains. The comparable effects for individuals are of the same sign but are not statistically significantly different to each other. Team productivity is enhanced by 1% purely due to the framing manipulation. Neither the framing nor the incentive effect lose their importance over the six month study period. Nor are there any detrimental effects on the quality of work as measured by product defect rates.

On a practical note, the results highlight that conditional on bonuses being provided, framing matters, and as framing can be adjusted almost costlessly, there are simply ways in which firms can further enhance productivity responses to monetary incentives. Theoretically, these results from the field provide an example of the prevalence of loss aversion in a natural labor market setting. As such the results provide external validity to laboratory evidence, and should be seen to provide a strong argument in favor of field and laboratory experiments being complements, not substitutes.

4.2. Non-monetary incentives

Organizations use a variety of non-pecuniary based incentives to motivate their employees. We discuss three forms of non-monetary incentive: status goods, feedback, and social incentives.[51]

[51] This list is not meant to be exhaustive. We focus on these because field experiments have provided insights on these margins to a greater extent than for other types of non-monetary incentive such as those discussed in Francois (2000), Dixit (2002), Prendergast (2001), Benabou and Tirole (2003), Seabright (2002), Delfgaauw and Dur (2004), Akerlof and Kranton (2005), and Besley and Ghatak (2005).

Under status incentive schemes, employees are given some positional good, such as an "employee of the month" job title. The notion that individuals crave status has been long studied (Veblen, 1934; Friedman and Savage, 1948; Duesenberry, 1949; Frank, 1985) and more recently formalized in the context of organizations providing status incentives in Moldovanu et al. (2007) and Besley and Ghatak (2008). They emphasize that for status incentives to be effective, the positional good must be valued by employees, it must be scarce, and its allocation rule rewards the deserving.

Recent evidence on these effects have been found in laboratory settings (Ball et al., 2001; Brown-Kruse et al., 2007) but few field experiments in which researchers have worked closely with a firm to exogenously vary such status rewards. One exception is Greenberg (1988), who reports results from a field experiment based on 198 employees in the underwriting department of a large insurance company. These employees were randomly assigned on a temporary basis to the offices of either higher, lower, or equal-status co-workers while their own offices were being refurbished. Relative to those workers reassigned to equal-status offices, those reassigned to higher status offices raised their performance, and those reassigned to lower status offices lowered their performance. The size of these performance changes were directly related to the magnitude of the status changes encountered. The results are interpreted as providing real world evidence on equity theory for non-monetary rewards.

In the future, we envisage field experiments being designed that randomly vary the first two margins on effective status incentives described above: how valued the positional good is, and its scarcity. In contrast, a field experiment that randomly allocated such positional goods might not be as informative, unless it was clearly related to some reallocation that would have occurred in any case, such as in Greenberg's clever study described above. Otherwise, such random allocations would not be representative of the kinds of allocation rule employers actually use, and so cloud the interpretation of any such results.[52]

A second class of non-monetary incentive in organizations relates to the provision of feedback. While there is a long tradition in psychology on feedback effects (Thorndike, 1913), economists have only recently begun to investigate its causes and consequences. Much of this research has focused on the theory of optimal feedback provision as mid-term reviews (Lizzeri et al., 2002; Ederer, 2008). Theory indicates that feedback on past performance can affect current performance either directly if past and current performances are substitutes or complements in the agent's utility function, labelled a

[52] There are field experiments on charitable giving that have exogenously varied the visibility of donations to assess whether such status concerns or prestige motives drive giving behavior Soetevent (2005), something that has been found to be the case in laboratory settings of public goods games (Andreoni and Petrie, 2004; Rege and Telle, 2004). Echoing some of the results below on gift-exchange in the field and the lab, Soetevent (2005) finds evidence that for some times of charitable cause, contributions increase when they can be socially recognized, but that this effect diminishes over time.

preference effect, or indirectly by revealing information on the marginal return to current effort, labelled a *signaling* effect.[53] The direct preference effect is relevant if, for instance, agents are compensated according to a performance target or fixed bonus scheme, so that being informed of high levels of past performance induces the agent to reduce her current effort relative to her past effort, and still meet her overall performance target. The indirect signaling effect is relevant if, for instance, the agent's marginal return to effort depends on her ability and this would be unknown if feedback were not provided.[54]

Both these mechanisms imply the effect of feedback is heterogeneous across individuals and it might increase or reduce current effort, so the socially optimal provision of feedback remains ambiguous. Research in psychology also suggests feedback effects are heterogeneous and may crowd in or crowd out intrinsic motivation (Butler, 1987; Deci et al., 2001). Consistent with such heterogeneous effects, the organizational behavior literature finds that performance feedback within firms is far from ubiquitous (Meyer et al., 1965; Beer, 1990; Gibbs, 1991).[55]

While there is a growing empirical literature on the effect of feedback in laboratory settings (Eriksson et al., 2008a,b), and in natural experiments (Bandiera et al., 2009a,b; Azmat and Iriberri, 2009; Blanes-i-Vidal and Nossol, 2009), evidence from field settings remains scarce. Bandiera et al. (2009a,b) provide one such analysis, in which the focus is on the provision of feedback to teams. This study is described in detail below.[56]

The third class of non-monetary incentives relate to changes in behavior induced because of the presence and identity of co-workers—namely social relations in the workplace. The idea that there exists an interplay between social relations and monetary incentives in the workplace goes back to the old Hawthorne studies mentioned earlier and have been long considered in the organizational and business sociology literatures (Mayo, 1933; Barnard, 1938; Roethlisberger and Dickson, 1939; Roy, 1952;

53 The organizational behavior and psychology literatures have also emphasized the signaling effects of feedback, as well as other related comparative statics such as how individuals change strategies in response to feedback (Vollmeyer and Rheinberg, 2005), and the specific type of information that should be conveyed in feedback (Butler, 1987; Cameron and Pierce, 1994).

54 Two strands of the economics literature have explored aspects of the signaling effect of feedback. The first strand focuses on whether individuals update their priors in response to feedback consistent with Bayes' rule (Slovic and Lichtenstein, 1971). The second strand focuses on whether agents react more to positive than negative feedback because of self serving biases such as confirmatory bias (Rabin and Schrag, 1999), or overconfidence (Malmendier and Tate, 2005; Van Den, 2004).

55 On the heterogeneous effects of feedback, the meta-analysis of Kluger and Denisi (1996) covering 131 studies in psychology with 13,000 subjects finds that two thirds of studies report positive feedback effects. On the optimal provision of feedback, when the agent knows her ability so that there is no indirect signaling effect of feedback, whether feedback should be optimally provided or not is sensitive to the specification of the agent's cost of effort function (Lizzeri et al., 2002; Aoyagi, 2007). More general results have been derived when agents learn their ability through feedback and ability is complementary to effort (Ederer, 2008).

56 A separate branch of the literature has focused on the strategic manipulation of feedback by the principal (Malcolmson, 1984; Gibbs, 1991; Aoyagi, 2007), of which there is anecdotal evidence from the field (Longnecker et al., 1987) and laboratory (Ederer and Fehr, 2007). Evidence from field experiments on feedback remains scarce.

Williams and O'Reilly, 1998). Such concerns have begun to be incorporated in economic theory (Kandel and Lazear, 1992; Rotemberg, 1994; Prendergast and Topel, 1996), and credible evidence on their existence, magnitude and underlying mechanisms have begun to emerge in a nascent literature using non-experimental methods in combination with personnel data (Mas and Moretti, 2009; Bandiera et al., 2010, forthcoming).

In a series of natural field experiments, Bandiera et al. provide evidence on the effect of incentives on individual and firm performance within the same firm. These field experiments engineer exogenously timed variation in the incentive structures faced by workers in the firm. The common thread running through these studies is to provide evidence on the interplay between monetary and non-monetary incentives in the workplace. The specific form of non-monetary incentives considered are those arising from social relations in the workplace, so that workers behavior, and response to monetary incentives, might differ depending on the nature of the social ties they have with co-workers, their superiors, and their subordinates. Given that this form of non-monetary incentive is what field experiments have predominantly focused on, we first develop a framework that makes precise how such incentives can be incorporated into an otherwise standard model, and then map this framework to the empirical evidence from the field.

4.2.1. Theoretical framework

Suppose worker i's payoff depends on three components. First, she derives some benefit from exerting effort e_i towards a productive task. This benefit, $B(e_i, .)$, reflects in part how her effort maps into income through the monetary compensation scheme. To cover a wide range of compensation schemes including absolute performance evaluation incentive schemes such as piece rates, relative performance evaluation schemes such as rank order tournaments, or team incentives, these benefits will in general also depend on co-workers' effort, $\mathbf{e}_{-i}$. Second, the worker faces a convex cost of effort, $C(\theta_i, e_i)$, where workers are of heterogeneous ability, θ_i. Finally, we assume worker i places some weight on the utility of co-worker j, π_{ij}. In turn, such social preferences π_{ij} might depend on the existence or strength of the social tie between individuals i and j. This third component of the worker's payoff function generates social incentives.[57] Workers simultaneously choose their efforts to maximize their total payoff,

$$\max_{e_i} B_i(e_i, \mathbf{e}_{-i}) - C(\theta_i, e_i) + \sum_{j \neq i} \pi_{ij}[B_j(e_j, \mathbf{e}_{-j}) - C_j(\theta_j, e_j)]. \tag{29}$$

[57] Social preferences can be thought of as a reduced form representation of a number of models. They depict behavior consistent with reciprocity or altruism (Fehr and Schmidt, 1999), or the evolutionary equilibrium of a repeated Prisoner's Dilemma game in which workers learn which strategies to play (Levine and Pesendorfer, 2002; Sethi and Somanathan, 1999). In the field experiment reported in Bandiera et al. (2005), they attempt to distinguish between models in which workers' preferences display altruism towards others, and models in which workers behave *as if* they are altruistic because, for instance, they play trigger strategies to enforce implicit collusive agreements.

The first order condition is[58],

$$\frac{\partial B_i(.)}{\partial e_i} - \frac{\partial C(\theta_i, e_i)}{\partial e_i} + \sum_{j \neq i} \pi_{ij} \frac{\partial B_j(.)}{\partial e_i}. \tag{30}$$

The monetary compensation scheme determines the marginal benefit of effort, $\frac{\partial B_i(.)}{\partial e_i} \geq 0$. As the worker has social incentives, she takes account of the fact that on the margin, her effort also affects the benefits that accrue to others, $\frac{\partial B_j(.)}{\partial e_i} \lesseqgtr 0$. As mentioned above, the precise sign of this social interaction depends on the nature of peer effects between workers that are socially connected, and the monetary compensation scheme in place.

The theoretical predictions of such models generate a wide range of behavioral responses. For example, working alongside friends might make work more enjoyable, generate contagious enthusiasm among friends, provide positive role models, or generate incentives to compete to be the best in the network of friends. All such mechanisms, that effectively increase the net benefits of effort, imply workers exert more effort in the presence of their friends relative to themselves when they work in the absence of their friends. Alternatively, working with friends might create contagious malaise, or lead to low effort norms within friends or co-workers more generally. All such mechanisms, that effectively decrease the net benefits of effort, imply workers exert less effort in the presence of their friends. Finally, the presence of friends might have heterogeneous effects across workers in that some exert more effort in the presence of their friends relative to when they work solely with non-friends, and others exert less effort. For example, friends or co-workers may conform to a common norm (Bernheim, 1994), or workers might be averse to pay inequality within their network (Fehr and Schmidt, 1999; Charness and Rabin, 2002). In either case, relative to when they work only with non-friends— (i) low ability workers exert more effort in the presence of their friends, and; (ii) high ability workers exert less effort in the presence of their friends. These aspects are highlighted by the field experiments discussed below on non-monetary incentives.

The field experiments in Bandiera et al. are designed to engineer exogenous variation in the incentives faced by workers to identify $\frac{\partial B_i(.)}{\partial e_i}$, corresponding to a similar reduced form parameter as in Lazear (2000) and Shearer (2004). They then combine this variation with primary data collected on social networks and plausibly exogenous variation in the assignment of friends as co-workers over time, to identify social incentives as embodied in $\frac{\partial B_j(.)}{\partial e_i}$. In these experiments, the authors examine the effects of social incentives both within and across tiers of the firm hierarchy. Namely in some studies i and j are co-workers engaged in the same tasks, and in other studies the pair correspond to a manager

[58] The model would be complicated if there were also knowledge spillovers such that effort exerted by worker i reduced the cost of effort of worker j. While such knowledge spillovers have been found in workplace settings (Moretti, 2004; Ichniowski et al., 1997) we abstract from them here.

and her subordinate. Moreover, they study cases in which: (i) individual effort hurts co-workers $\frac{\partial B_j(.)}{\partial e_i} < 0$, as in the case of relative incentive schemes; (ii) where it benefits them $\frac{\partial B_j(.)}{\partial e_i} > 0$, as in the case of team incentives; and (iii) where it has no effect $\frac{\partial B_j(.)}{\partial e_i} = 0$, as in the case of a piece rate scheme. We now summarize the main insights from these natural field experiments.

4.2.2. Evidence from the field

Social incentives among bottom tier workers

The firm studied in Bandiera et al. is a leading UK producer of soft fruit. Managerial staff belongs to three classes. The first class consists of a single general manager whom we refer to as the Chief Operating Officer (COO), the second comprises ten field managers, and the bottom-tier of the firm hierarchy consists of workers whose main task is to pick fruit. Field managers are responsible for field logistics, most importantly to assign workers to rows of fruit within the field and to monitor workers. Managerial effort can therefore be targeted to individual workers and is complementary to worker's effort. The main task of the COO is to decide which workers are selected to pick fruit each day, and which are assigned to non-picking tasks. The field experiments described below together provide insights on behavior at each tier of the firm's hierarchy.

In each natural field experiment, the researchers worked closely with the CEO of the firm to engineer exogenously timed changes in monetary incentives to workers or managers. The *same* workers and managers are observed under both incentive schemes and therefore it is possible to control for time invariant sources of heterogeneity across workers, such as their ability, and across managers, such as their management style.[59] The most important remaining empirical concern is that the estimates of such changes might still reflect naturally occurring time trends in productivity. This is addressed using a battery of tests in each paper. In addition, the time span of study allows the authors to check in each case whether the behavioral response to incentives is long-lasting, or whether they reflect Hawthorne effects, as discussed earlier, whereby individuals respond in the short run to any change in their workplace environment. Being able to use field experiments to estimate short and long run responses to changes in management practice is a theme we will return to below when we present field experimental evidence on gift-exchange in firms, and contrast the evidence from the field and the laboratory.[60]

[59] Hence this empirical strategy is informed by the evidence that individual "styles" of managers affect firm performance over and above firm level characteristics themselves (Bertrand and Schoar, 2003; Malmendier and Tate, 2005).

[60] Bandiera et al. study the behavior of nearly all the workers in the firm for each field experiment. However, given the experiment takes place in one firm, to avoid contamination effects across treated and control groups, all workers were simultaneously shifted from one incentive scheme to the other. In contrast, Shearer (2004) exogenously varied the incentive scheme workers were in on each day. In non-experimental studies such as Lazear (2000) on individual pay and Hamilton et al. (2003) on team pay, workers might have had some say on which compensation scheme they would be paid under.

In each natural field experiment, the authors collected primary data on the social networks of each individual worker. With such a precise mapping of the structure of friendship networks in the firm, personnel data providing workers productivity over time, and the field experiment on monetary incentives, the authors are able to shed light on the interplay between monetary and social incentives in this setting.

Finally, they have daily information on the pool of workers available to pick fruit. This allows them to precisely identify the effect of monetary incentives on the selection of workers from this pool. The entire pool of workers is observed in this context because individuals are hired seasonally from Eastern Europe, and they live on the farm for the duration of their stay. This margin of selection—driven by the COO's *demand* for workers—from the firm's internal labor market proves to be an important margin of response to some changes in incentives, particularly in relation to changes in managerial incentives. Still, these field experiments, like Shearer (2004), are silent on the selection effect highlighted by Lazear (2000) in relation to workers choice of which firm to *supply* their labor to.

Another obvious similarity between Shearer (2004) and Bandiera et al. is that they study agricultural environments in which worker productivity is easy to measure, comparable across workers at the same moment in time, and comparable within a worker over time. The fact that worker productivity is measured electronically with little measurement error, also makes analysis of the impact of the field experiment on the distribution of productivity, again as highlighted by Lazear (2000), particularly amenable to quantile regression methods for example. However, it remains true that settings in which worker's output is hard to measure, verify or compare, which might represent the bulk of tasks in the modern service based economy, remain relatively unexplored in field experiments.

In Bandiera et al. (2005) the natural field experiment exogenously changes the monetary incentives to the bottom-tier workers whose primary task is to pick fruit. The study compares the behavior of these workers under a relative incentive scheme to a piece rate scheme. The comparison is revealing because under relative incentives individual effort imposes a negative externality on co-workers' pay whereas under piece rates individual effort has no effect on others' pay. The difference in workers' performance under the two schemes, if any, then provides evidence on whether and to what extent workers internalize the externality they impose on their colleagues. To see this, the framework above is tailored to this specific field experiment as follows.

Consider a group of N workers, each worker i exerts effort $e_i \geq 0$ which determines her productivity. The cost of effort is assumed to be $\frac{\theta_i e_i^2}{2}$. Under relative incentives the benefit from pay depends on the worker's productivity relative to all her co-workers, $B\left(\frac{e_i}{\bar{e}}\right)$, where $\bar{e} = \frac{1}{N}\sum_i e_i$. The relative scheme has the key characteristics that an increase in worker i's effort—(i) increases her pay; (ii) increases average effort and hence imposes a negative externality by reducing the pay of co-workers. The effort choice under relative incentives then depends on whether workers have social incentives and

therefore internalize this externality. Assuming worker i places the same social weight on all co-workers, so $\pi_{ij} = \pi_i$, the equilibrium effort for worker i solves,

$$\max_{e_i} B\left(\frac{e_i}{\overline{e}}\right) + \pi_i \sum_{j \neq i} \left(B\left(\frac{e_j}{\overline{e}}\right) - \frac{\theta_j e_j^2}{2} \right) - \frac{\theta_i e_i^2}{2}. \tag{31}$$

Assuming worker i chooses her effort taking the effort of others as given, the Nash equilibrium effort for worker i solves,

$$B'\left(\frac{e_i}{\overline{e}}\right)\frac{1}{\overline{e}}\left(\frac{\sum_{j \neq i} e_j}{\left(\sum_i e_i\right)}\right) - \frac{\pi_i}{\overline{e}} \sum_{j \neq i} B\left(\frac{e_j}{\overline{e}}\right)\frac{e_j}{\left(\sum_i e_i\right)} = \theta_i e_i. \tag{32}$$

Under piece rates, individual effort is paid at a fixed rate b per unit and worker i chooses her effort as follows,

$$\max_{e_i} B\left(be_i\right) + \pi_i \sum_{j \neq i} \left(B\left(be_j\right) - \frac{\theta_j e_j^2}{2} \right) - \frac{\theta_i e_i^2}{2}. \tag{33}$$

The equilibrium effort level solves the first order condition,

$$B'\left(be_i\right) b = \theta_i e_i. \tag{34}$$

As worker i's effort does not affect her co-workers' pay, her optimal choice of effort is independent of π_i. To compare effort choices under the two schemes, evaluate (34) at $b = \frac{1}{\overline{e}}$ so that for a given $\overline{e}$, the pay per unit of effort is the same under both incentive schemes. The first order condition under piece rates then is,

$$B'\left(\frac{e_i}{\overline{e}}\right)\frac{1}{\overline{e}} = \theta_i e_i, \tag{35}$$

so the difference between the first order conditions (32) and (35) can be ascribed to two sources. The first is the externality worker i imposes on others under relative incentives, the magnitude of which depends on π_i. When $\pi_i > 0$ worker i's productivity is *lower* under relative incentives compared to piece rates. Second, by exerting more effort, each worker lowers the pay she receives for each unit of effort under relative incentives. This effect, captured by the $\frac{\sum_{j \neq i} e_j}{\left(\sum_i e_i\right)}$ term, also reduces productivity under relative incentives but is negligible in large groups.

The main results from Bandiera et al. (2005) are then as follows. First, the reduced form estimates suggest that the exogenously timed switch from relative incentives to piece rates had a significant and permanent impact on worker productivity. For the average worker, productivity increased by at least 50% moving from relative incentives to piece rates. As in the earlier literature, both the mean and dispersion of productivity significantly increase with the move to piece rates. The productivity gains achieved under piece rates are not found to be at the expense of a lower quality of picking.

The authors then assess whether this productivity change is consistent with the standard assumption that workers ignore the externality they impose on others under the relative scheme ($\pi_i = 0$), or whether they fully internalize it ($\pi_i = 1$). To do this they use the structural model above, imposing a functional form assumption on $B(.)$ and a production function linking effort to observed output, to calibrate the first order conditions of the workers' maximization problem to compute an estimate of each worker's cost parameter, θ_i, under each incentive scheme and behavioral assumption. Since worker's ability is innate, they ought to find the *same* implied distributions of costs across workers under both incentive schemes if the underlying behavioral assumption is correct.

Calibration of the first order conditions for worker's efforts reveals that the observed change in productivity is *too large* to be consistent with the assumption that workers ignore the negative externality they impose on others. At the same time, the observed change in productivity is also *too small* to be consistent with the assumption that workers maximize the welfare of the group and fully internalize the negative externality. The authors then uncover the distribution of social weights π_i across workers that would explain the productivity increases. To do so they assume the true cost of effort θ_i of each worker is that derived under piece rates, and then substitute into the first order condition (32). They find the data is consistent with the average worker placing a weight of $\overline{\pi} = 0.65$ on the benefits accruing to all other co-workers, assuming they place a weight of one on their own benefits.

Further analysis combines the experimental variation induced by the change in incentive scheme, with non-experimental variation of the assignment of workers to work alongside their friends on some days but not on other days. The field experiment method allows the collection of primary data on social networks of each worker on the farm. This reveals that under relative incentives workers internalize the externality more when the share of their personal friends in the group is larger and this effect is stronger in smaller groups. In line with the interpretation that social preferences explain the difference in productivity across the two schemes, the relationship among workers *does not* affect productivity under piece rates. Finally, they find that productivity under relative incentives was significantly lower only when workers were able to monitor each other. Given that monitoring is necessary to enforce collusion while it does not affect altruism, they take this finding to support the hypothesis that workers are able to sustain implicit collusive agreements when relative incentives are in place. Hence, building on

a large body of evidence from laboratory settings, this evidence from the field suggests workers behave as if they have social preferences but do not, in structural form, have social preferences that make them unconditionally altruistic towards others.

The results beg the question of why, given the large gains to productivity and profits, of the move to piece rates, were relative incentives ever employed in the first place. The farm management suggested the relative scheme was mainly adopted to difference out common shocks that are a key determinant of workers productivity in this setting. While this is in line with the predictions of incentive theory, the superiority of relative incentives relies on the assumption that workers ignore the externality their effort imposes on others.[61] This assumption on worker behavior is not supported by this field experiment. Relative incentives led to lower productivity because workers internalized the negative externality to some extent. The results of this natural field experiment then speak directly to Lazear's (1989) observation on how rarely workers are compensated according to rank-order tournaments, and point to new and interesting directions for theory to develop on the optimal provision of incentives under more robust assumptions on worker preferences.

Social incentives among managers

While the evidence from field experiments discussed thus far has focused on the monetary incentives provided to bottom-tier workers, Rosen's (1982) magnification principle implies the incentives provided higher up in the firm hierarchy can have larger effects on firms' performance. Bandiera et al. (2007) present evidence from a field experiment in the same setting as previously described to explore this issue.

They examine the effects of providing bonuses to managers based on the average productivity of their subordinates. They extend the framework above to highlight that, as in most firms, in their context managers can affect worker productivity through two channels—(i) they can take actions that affect the productivity of existing workers, and, (ii) they can affect the identity of the workers selected into employment. A simple theoretical framework indicates that, when workers are of heterogeneous ability and managers' and workers' effort are complements, the introduction of managerial performance pay makes managers target their effort towards the most able workers. This is labeled a "targeting effect" of managerial incentives. In addition, the introduction of managerial performance pay makes managers select the most able workers into employment. This is labeled as a "selection effect" of managerial incentives.

As in Lazear's framework, such targeting and selection effects influence both the mean and the dispersion of workers' productivity. Mean productivity unambiguously rises as managers target the most able workers and fire the least able. The effect on the

[61] See Lazear and Rosen (1981), Green and Stokey (1983) and Nalebuff and Stiglitz (1983). Relative performance evaluation may also be preferred to piece rates as it lowers informational rents to high types (Bhaskar, 2002), and reduces incentives of workers to exert effort in influence activity (Milgrom, 1988).

dispersion is however ambiguous. On the one hand, targeting the most able workers exacerbates the natural differences in ability and leads to an increase in dispersion. On the other hand, if only more able and hence more similar workers are selected into employment in the first place, the dispersion of productivity may fall, depending on the underlying distribution of ability across workers.

They key findings from Bandiera et al. (2007) are as follows. First, the introduction of managerial performance pay increases both the average productivity and the dispersion of productivity among lower-tier workers. The average productivity increases by 21 percent and the coefficient of variation increases by 38 percent.

Second, the increase in the mean and dispersion of productivity is due to both targeting and selection effects. The analysis of individual productivity data reveals that the most able workers experience a significant increase in productivity while the productivity of other workers is not affected or even decreases. This suggests that the targeting effect is at play—after the introduction of performance pay, managers target their effort towards more able workers. The individual data also provides evidence of a selection effect. More able workers, namely those who had the highest productivity when managers were paid fixed wages, are more likely to be selected into the workforce when managers are paid performance bonuses. Least able workers are employed less often and workers at the bottom of the productivity distribution are fired.[62]

Third, the selection and targeting effect reinforce each other, as workers who experience the highest increase in productivity are also more likely to be selected into employment. The introduction of managerial performance pay thus exacerbates earnings inequality due to underlying differences in ability both because the most able workers experience a larger increase in productivity and because they are selected into employment more often.

Finally, they evaluate the relative importance of the targeting and selection effects through a series of thought experiments. They find that at least half of the 21 percent increase in average productivity is driven by the selection of more productive workers. In contrast, the change in dispersion is nearly entirely due to managers targeting the most able workers after the introduction of performance pay. Namely, the dispersion of productivity would have increased by almost the same amount had the selection of

62 The results from this natural field experiment has implications for environments outside the workplace. For example, the provision of teacher incentives based on the average performance of students may have important consequences for the distribution of test scores among students, and the composition of students, and possibly teachers, admitted into schools. For example, Burgess et al. (2005) find that the introduction of school accountability based on test pass rates improved the performance of students in the middle of the ability distribution, at the expense of both high achieving and low achieving students. Similarly, Hanushek and Raymond (2004) and Reback (2005) provide evidence on the distributional consequences on student achievement under the *No Child Left Behind* policy. Finally, Jacob (2002) and Figlio and Getzler (2002) provide evidence on the selection effect. They show that the introduction of accountability schemes lead to an increase in grade retention and special educational placement in Chicago and Florida public schools, respectively.

workers remained unchanged. The reason is that the distribution of ability across workers is such that even when the least able workers are fired, the marginal worker selected to pick is still of relatively low ability. Hence there remains considerable heterogeneity in productivity among selected workers.

These findings shed some light on why firms provide performance related pay to managers in the first place. While such incentive schemes are obviously designed to increase unobservable managerial effort, these results suggest another more subtle reason for their use. This stems from the general observation that firms are typically constrained to offer bottom-tier workers the *same* compensation scheme. This may be because of legal, technological or informational constraints (Lazear, 1989; Bewley, 1999; Encinosa et al., 1997; Fehr et al., 2004). To the extent that bottom-tier workers are of heterogeneous ability, however, offering the same compensation scheme to all of them will be sub-optimal. When managers' pay is linked to firm's performance, their interests become more aligned with those of the firm and they have greater incentives to target their effort to specific workers in order to offset the inefficiency that arises because of the common compensation scheme. From the worker's point of view it is then as if they face an individual specific incentive scheme. This opens a broad research agenda to examine whether firms are indeed more likely to offer managers performance pay in settings where lower tier workers are of heterogeneous ability, managers are able to target their effort towards specific workers, and workers are offered the same compensation scheme.

The findings from this field experiment also highlight the interplay between the provision of managerial incentives and the earnings inequality among lower-tier workers. Such a linkage exists whenever managers can target their efforts towards some workers and away from others, and managers choose which individuals are selected into the workforce. Hence that there might be an important interplay between managerial incentives and earnings inequality among workers highlights a possible link between two important trends in labor markets over the past twenty years that have previously been unconnected in the economics literature—the rising use of managerial performance pay, and the rising earnings inequality among observationally similar workers.[63]

In Bandiera et al. (2009a,b), the authors use the same introduction of managerial bonuses to understand whether managers favor workers they are socially connected to. In general, social connections between managers and workers can help or harm firm performance. On the one hand, social connections may be beneficial to firm performance if they allow managers to provide non-monetary incentives to workers, or help reduce informational asymmetries within the firm. On the other hand, managers may

63 Residual, or within-group wage inequality, is a sizeable contributor of the growth in overall wage inequality in the US. This has been argued to have increased throughout the 1970s and 1980s (Juhn et al., 1993), and into the 1990s (Acemoglu, 2002; Autor et al., 2005).

display favoritism towards workers they are socially connected with, to the detriment of other workers and overall firm performance.[64]

In this experiment, as managerial compensation becomes more closely tied to firm performance, we would expect managers to utilize social connections to a greater extent if indeed, such connections are beneficial for firm performance. On the other hand, if social connection are bad for the firm, we might expect managers to reallocate their effort across workers in response to managerial incentives, towards high ability workers, and away from workers they are socially connected to. To be precise, if the managers' behavior towards connected workers changes once their interests are more closely aligned with the firm's, their previous behavior under fixed wages could have not been maximizing the firm's average productivity.

To measure social connections the authors use a survey they designed to exploit three sources of similarity between managers and workers—whether they are of the same nationality, whether they live in close proximity to each other on the farm, and whether they arrived at a similar time on the farm. The underlying assumption is that individuals are more likely to befriend others if they are of the same nationality, if they are neighbors, or if they share early experiences in a new workplace.[65]

The main findings are as follows. First, when managers are paid fixed wages, the productivity of a given worker is 9% higher when he is socially connected to his manager, relative to when he is not. As workers are paid piece rates, this translates into the same proportionate change in earnings. Second, when managers are paid performance bonuses that tie their pay to the average productivity of workers they manage, being socially connected to the manager has no effect on workers' productivity.

Third, the introduction of managerial performance pay significantly decreases the productivity of low ability workers when they are connected to their manager relative to when they were connected to their manager and she was paid a fixed wage. The introduction of managerial performance pay increases the productivity of high ability workers, especially when they are not connected to their managers. These findings indicate that when managers face low powered incentives, they favor the workers they are socially connected to, regardless of the workers' ability. In contrast, when they face high powered incentives, managers favor high ability workers regardless of the workers' connection status.

64 Both the positive and negative effects of social connections have been stressed in the organizational behavior and sociology literatures. Examples of such work includes that on the effect of manager-subordinate similarity on subjective outcomes such as performance evaluations, role ambiguity, and job satisfaction (Tsui and O'Reilly, 1989; Thomas, 1990; Wesolowski and Mossholder, 1997), and on how social networks within the firm influence within firm promotions (Podolny and Baron, 1997).

65 Lazear (1989), Kandel and Lazear (1992), and Rotemberg (1994) develop models incorporating social concerns into the analysis of behavior within firms. While they emphasize that individuals have social concerns for others at the same tier of the firm hierarchy, their analysis is equally applicable across tiers of the hierarchy. Bewley (1999) offers extensive evidence from interviews with managers arguing that concerns over fair outcomes for workers and the morale of employees are important determinants of their behavior.

Fourth, an increase in the level of social connections between managers and workers has a detrimental effect on the firms' average productivity when managers are paid fixed wages and has no effect when managers are paid performance bonuses. In this setting, social connections are therefore detrimental for the firm because their existence distorts the allocation of managerial effort in favor of lower ability workers.

This natural field experiment paper contributes to the growing empirical evidence on the interplay between social networks and individual and firm performance. In particular, the design allows the authors to identify not only whether social connections matter within the firm, but also exploit the exogenous variation in incentives to understand whether they are to the benefit or detriment of the firm.

Feedback

In a final natural field experiment from this setting, Bandiera et al. (2010, forthcoming) present evidence to evaluate the effect of performance feedback and monetary prize tournaments, when the workforce is organized in teams. Hence in this set-up workers effort imposes a positive externality on their team members, $\frac{\partial B_j(.)}{\partial e_i} > 0$. They compare the effects of these forms of non-monetary and monetary incentives relative to when teams are paid piece rates, and analyze their effect on two outcomes: how workers sort into teams and team productivity.

This field experiment provides important contributions to the literature along three margins. First, despite the pervasiveness of teams in the workplace, field evidence on team incentives is scarce.[66] The existing evidence from individual reward schemes provides limited guidance because the margins along which individuals and teams can respond to incentives differ. Specifically, in addition to changes in individual effort, changes in team incentives can lead to changes in team composition. To the extent that workers effort depends on the identity of their team members because of social incentives, changes in team composition can affect the productivity of the individual teams and of the firm as a whole.[67]

Second, tournaments are widely used to provide incentives across diverse organizations such as salespeople competing for bonuses, managers competing for promotions, and politicians competing for vote shares (Bull et al., 1987; Baker et al., 1988). While several studies have tested whether the response to variation in tournament structure is consistent with theoretical predictions, field evidence on the comparison of

[66] More than 70% of major US firms use some form of team based rewards (Ledford et al., 1995). Lazear and Shaw (2007) cite evidence that between 1987 and 1996, the share of large firms that have more than a fifth of their employees in problem solving teams rose from 37 to 66%. The percentage of large firms with workers in self-managed teams rose from 27 to 78% over the same period. In academia, Wuchty et al. (2007) document the increased use of team production in research across disciplines.

[67] There is only a small literature on selection into teams in laboratory settings (Weber, 2006; Charness and Yang, 2008), although there is a far more extensive lab-based literature on team production, as reviewed in Charness and Kuhn (2011, 2010).

monetary prize tournaments against alternative monetary and non-monetary incentive mechanisms is scarce.[68]

Third, whenever tournaments are in place, workers inevitably receive some information on their relative performance. This information might have direct effects on productivity if individuals have concerns for their relative position or status (Moldovanu et al., 2007; Besley and Ghatak, 2008), inequality aversion (Fehr and Schmidt, 1999; Charness and Rabin, 2002) or conformity (Bernheim, 1994). The field experiment allows the authors to de-couple the effect of feedback from the effect of monetary prize tournaments. As the provision of feedback is almost costless, measuring its contribution to the overall tournament effect can lead to considerable cost savings if most of the positive effect of tournaments on productivity is actually due to worker responses to feedback.[69]

In the experiment, at the beginning of the season, teams were paid piece rates based on their aggregate productivity. Halfway through the season teams were additionally provided feedback by posting daily histograms of each team's productivity. This feedback makes precise the absolute productivity of each team, and their ranking relative to all other teams. Halfway through the remaining part of the season a monetary prize for the most productive team each week was introduced, in addition to the provision of feedback, and conditional on teams being paid according to piece rates.

When workers first arrive at the farm they are assigned to a team by the general manager for their first week. Thereafter workers are free to choose their own team members at a *team exchange* that takes place every week. A team is formed only if all its members agree. Hence in this setting workers have two choice variables: how much effort to exert into picking, and team composition.

The field experiment is again closely tied to an underlying model. This makes precise two key forces that drive team formation: workers' ability and social connections. As individual earnings are increasing in the ability of team members, workers have incentives to assortatively match by ability. On the other hand, workers might prefer to form teams with friends because this might limit free-riding within teams (Alchian and Demsetz, 1972; Holmstrom, 1982; Kandel and Lazear, 1992), and because they enjoy

68 The empirical literature on tournament theory comprises two distinct branches. The first tests whether a particular compensation scheme has a tournament structure. Two specific predictions have been explored—(i) the wage spread should be positively related to the number of workers at the lower job level; (ii) the wage structure should be convex as in Rosen (1986). These tests typically use data from the market for CEOs (Gibbons and Murphy, 1990; Eriksson, 1999; Bognanno, 2001). The second branch of the literature tests whether individual behavior changes with tournament features in a way consistent with theory, using data either from experimental settings (Bull et al., 1987; Nalbantian and Schotter, 1997; Eriksson et al., 2008a,b; Freeman and Gelber, 2008), personnel data (Knoeber and Thurman, 1994; Eriksson, 1999; Bognanno, 2001), or sports (Ehrenberg and Bognanno, 1990). There are few existing field studies—on either individuals or teams—exploring tournament incentives to other incentive schemes such as piece rates or feedback.

69 Evidence from the laboratory has tended to focus on feedback to individuals (Freeman and Gelber, 2008). One exception is Sausgruber (2009) who provides experimental evidence on the effects on team performance when told about the performance of one other team, holding team composition constant.

non-pecuniary benefits from interacting with co-workers they are socially connected to Rosen (1986); Hamilton et al. (2003).[70] To the extent that workers are not socially connected to colleagues of similar ability, a trade-off emerges. The theoretical framework then makes precise how the introduction of feedback and prizes affect this trade-off.

The key empirical results from the field experiment are as follows. First, the introduction of feedback and of monetary prizes leads to significant changes in team composition. Relative to the piece rate regime, the share of team members connected by social ties is lower and team members' ability levels are more similar under the feedback and tournament regimes.

Second, the feedback and tournament schemes have opposite effects on average productivity. Relative to the piece rate regime, the introduction of feedback significantly reduces average team productivity by 14%. The further introduction of a monetary prize tournament, conditional on the provision of feedback, significantly increases productivity by 24%. As made precise in the theoretical framework, the reduction in average productivity when feedback is provided is consistent with workers being better off sorting into teams on the basis of ability rather than friendship as feedback increases the strength of incentives faced, and the firm being worse off because it no longer harnesses the ability of socially connected workers to ameliorate free-riding within the team. Hence the endogenous formation of teams under feedback reduces the firm's productivity overall. In contrast, the tournament incentives are sufficiently high-powered so the increase in worker's effort more than offsets any increase in free-riding within teams. Hence the firm's overall productivity rises.

Third, the dispersion of productivity increases under both regimes because both effects are heterogeneous as indicated by the theoretical framework. Quantile regression results show that the introduction of feedback reduces the productivity of teams at the bottom of the conditional productivity distribution compared to piece rates, while it has no effect on teams above the 40th percentile. In contrast, the introduction of prizes increases the productivity of teams at the top of the conditional productivity distribution compared to piece rates, while it has no effect on teams below the 30th percentile.

Fourth, focusing on the teams that remain intact after each change in incentives, the authors evaluate the effect of feedback and prizes on effort, holding constant team composition. They find that while the effect of feedback on team productivity is positive the magnitude appears small. This emphasizes that the documented negative effect of feedback is primarily due to the endogenous changes in team composition caused by the provision of feedback, rather than changes in behavior of the same team. In contrast

[70] In line with this, Rotemberg (1994) develops a model showing how altruism between co-workers may endogenously form in the workplace to facilitate cooperation among workers engaged in team production. Empirically, Hamilton et al. (2003) provide non-experimental evidence from the introduction of team production in a garment firm. They find the most able workers sorted first into teams despite a loss in earnings in many cases, suggesting non-pecuniary benefits associated with teamwork.

the additional introduction of monetary prizes increases team productivity by 25% for teams that choose to remain intact. Hence the provision of monetary prizes affects firm performance through both the endogenous changes in team composition and changes in behavior within the same team.

Finally, the authors present qualitative evidence from a worker survey they conducted. As highlighted at the start of this section, this type of primary data collection that is inherent in field experiments, allows the authors to shed light on other margins of behavior between workers that might be affected by the monetary and non-monetary incentives provided, but that the firm does not collect data on *ex ante*. This survey data reveals that relative to the piece rate regime, during the tournament regime significantly fewer workers report pushing their team members to work hard or giving team members instructions. This is consistent with workers being better matched by ability and having fewer social connections with their team members under the tournament regime, so that peer pressure within the team becomes less effective.

By exploring changes in behavior on a range of dimensions, this evidence from the field highlights new directions for research in understanding how agents react to monetary and non-monetary incentives in workplaces characterized by team production where teams form endogenously.

4.3. The employment relationship

The neoclassical labor market model emphasizes workers behave opportunistically. For example, in the model sketched above from Lazear (2000), when workers compensation is not tied to their performance, as under a fixed hourly wage scheme, all workers exert the minimum effort required to achieve the minimum output requirement, q_0. There is thus no variation in workers in their output or pay. We now explore the insights field experiments have provided on the existence and nature of such opportunistic behavior in real world settings. We do so through examples related to gift exchange in shirking.

4.3.1. Gift exchange

The standard labor market model assumes in equilibrium firms pay market clearing wages and workers provide minimum effort. This prediction does not receive uniform support empirically. There are numerous cases where employers are observed paying above the market equilibrium wage Akerlof (1982), and where workers exert more than the minimum effort level, as we have already discussed in relation to Lazear (2000) and in many other studies on employee performance under fixed wages. This has led to the development of the gift-exchange model which is based on the assumption of their being a positive association between wages and worker effort (Akerlof, 1982; Akerlof and Yellen, 1988, 1990). In this class of model, employers offer higher than market clearing

wages, and workers are viewed to positively reciprocate by providing higher than the minimum required effort.

Clearly such theories are hard to test using non-experimental data: there might be a host of unobservable factors that create a correlation between wages and worker effort. Hence, there has been a large body of evidence established in laboratory settings on gift-exchange in firm settings, which began with Fehr et al. (1993). In this original study, they constructed a labor market equilibrium with excess labor supply so that the equilibrium wage was low. Employees also had no pecuniary incentive to raise the quality of their work above the minimum required level, so the best response of employers was to pay the low equilibrium wage. Contrary to the prediction, the majority of employers attempted to induce employees to invest greater effort by offering them higher than market-clearing wages. On average, this high wage was reciprocated by greater employee effort. Overall, it was profitable for employers to offer high wage contracts.

Gneezy and List (2006) use a natural field experiment then look for evidence of gift-exchange in similar real world environments in which equilibrium wages are low and workers earnings are not tied to their performance. In moving from the lab to the field, one important comparative static to evaluate is how behavior changes with the duration of the task. In other words, are the types of positive reciprocity observed by workers in the lab, a long run phenomena. The psychology literature provides two reasons why the duration of tasks might matter. First, there is the distinction between hot and cold decision making (Loewenstein and Schkade, 1999; Loewenstein, 2005). Second, there can be adaptation of behavior over time (Gilbert et al., 1998).

Two subject pools were utilized for the field experiments. In each a between subject design was used. The first field experiment recruited undergraduate students to participate in an effort to computerize the holdings of a small library at the university. The task was to enter data regarding the books into a computer database. In the no-Gift treatment, individuals were offered a flat wage of $12 per hour. In the Gift treatment, once the task was explained to participants, they were surprisingly paid $20 per hour rather then $12 per hour as advertised. In total 19 workers were hired for six hours each; 10 were randomly assigned to the no-Gift treatment. The second field experiment was part of a door to door fundraising drive to support a university research center. Fundraising solicitors were recruited. All solicitors were told they would be paid $10 per hour, and those in the Gift treatment were surprisingly told they would actually receive $20 per hour. In total 23 solicitors were employed over two days, with 10 being randomly assigned to the no-Gift treatment.[71]

[71] In all such experiments, it is important to design the set-up to be able to distinguish gift-exchange from the alternative explanation of why there should be a positive relationship between wages and effort—efficiency wages. This hypothesis postulates employers pay above market-clearing wages to motivate workers to increase their effort level so as to avoid being fired, which reduces employer monitoring (Katz, 1986). Hence in both field experiments subjects were made aware that this was a one-time recruitment opportunity.

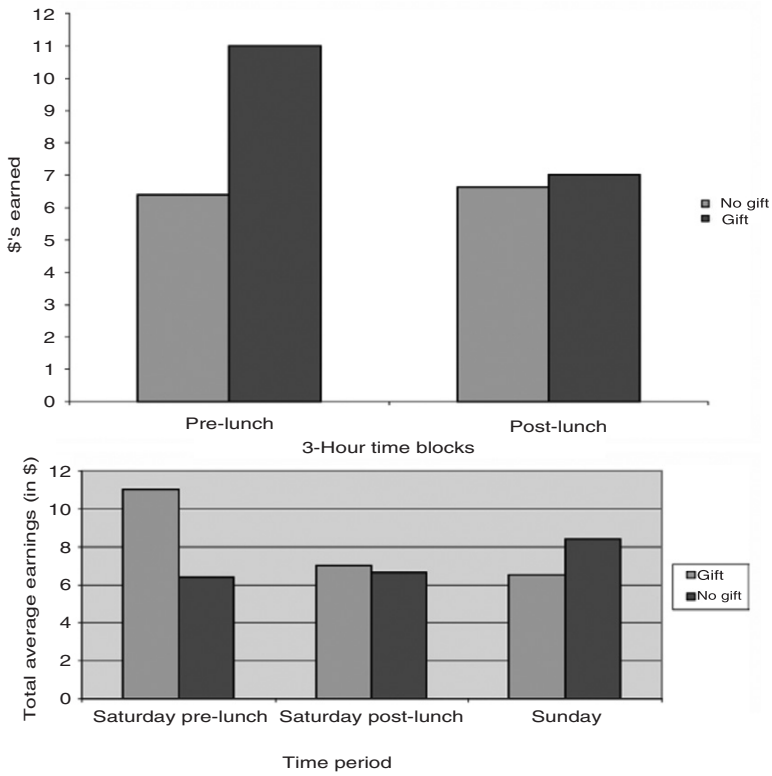


Figure 10 ***Gift exchange and the duration of tasks.***

The main results are as follows. First, in line with earlier evidence from laboratory settings, there are signs of significant gift exchange in the first few hours of the task, as measured both by effort in the library task and money raised in the fundraising task. For example, in the library task, effort is around 25% higher for those in the Gift treatment.

Second, there are significant falls in effort over time. After a few hours, there are no longer any significant differences in effort between the no-Gift and Gift treatments in either task. Figure 10 illustrates clearly how any positive reciprocity by workers is a short run phenomena in these two settings. Overall, the results suggest that with the same budget, the employer would have been better off just paying the market clearing wage as in the no-Gift treatments.

While the results go against the standard gift-exchange explanation of the positive association of wages and effort, they are in line with survey evidence on wage rigidity.[72]

[72] Clearly this debate on the existence of positive reciprocity in the field remains in need of further study. Charness and Kuhn (2011, 2010) describe more of the related evidence from the laboratory and other field settings.

For example Bewley (1999) considers why wages are downwardly rigid during a recession. He reports that managers are worried that wage cuts might result in decreases in morale that would subsequently result in poor worker performance when the economy recovered, if not immediately. This highlights the importance of fairness considerations in cases of negative reciprocity. With respect to positive reciprocity, as in this field setting, Bewley's evidence is less conclusive. He argues that morale is less important when considering wage increases, but finds that one main consideration when determining raises is the effect on employee turnover once the recession ends. Bewley's work suggests that there appears to be little connection between increasing pay and productivity, except to the extent that higher wages make it possible to attract, and retain, higher quality workers. This ties back to the earlier discussion of Lazear (2000) and the subsequent work on monetary incentives in firms, where it is thought there are qualitatively large selection effects of incentives driven by changes in workforce composition. Again, more evidence on these channels related to employee turnover are required to test more precisely a fuller set of theoretical predictions.

4.3.2. Shirking

Field experiments had also provided insights on research questions related to employee shirking behaviors. The standard economic framework emphasizes employees are rational shirkers: they will slack when the marginal benefits of doing so outweigh the marginal costs. Firms respond to such behaviors by choosing compensation and monitoring policies to reduce shirking. As emphasized above, this view that workers will behave so opportunistically when the marginal returns on their effort are low, is often contradicted by empirical evidence and predictions on behavior in such settings from the psychological and sociological literatures (Pfeffer, 1996; Kreps, 1997; Baron and Kreps, 1999). While we have earlier studied the role of compensation schemes and wage setting behavior to raise employee effort, we now focus on the effect employer monitoring has on worker behavior.

If employees are rational cheats then, conditional on a given incentive pay arrangement, a reduction in monitoring will lead to an increase in shirking. The most powerful sanction available to employers is typically dismissal. Thus, an increase in shirking resulting from reduced monitoring should be greatest among individuals for whom the ongoing employment relationship is least valuable.

As with most of the research questions posed in this chapter, establishing credible empirical evidence to support the theory is not straightforward. In this case there are two concerns that have plagued the non-experimental literature. First, shirking behavior is by its nature hard to detect. Moreover, the ability of the econometrician to detect shirkers might itself be endogenously related to the employer's monitoring practices. Second, there might be unobserved factors, such as other hiring policies, that cause there to be a correlation between monitoring and shirking.

A carefully designed natural field experiment of Nagin et al. (2002) addresses both challenges. The setting was a telephone solicitation company, with employees dispersed across 16 call centers. At each call center, telephone solicitors were paid according to the same piece rate incentive scheme, one in which salary increased with the number of successful solicitations. This piece rate, together with imperfect information on the outcome of pledges, created incentives for employees to falsely claim that they had solicited a donation.

To curb opportunistic behavior, the employer monitored for false donations by calling back a fraction of those who had responded positively to a solicitation. Employees were informed when hired that their activities would be checked by "callbacks" made by management. The results of each week's callbacks were communicated to both employees and their immediate supervisors, and the bad calls were deducted from each individual's weekly incentive pay. Stronger sanctions for bad calls were not generally imposed on employees because the number of bad calls was understood to be a noisy indicator of cheating. For example, donors might sometimes change their mind after agreeing to pledge money.[73]

To see if the costs of implementing this monitoring system could be reduced, the company conducted a controlled field experiment. This experiment was "double blind" in the sense that neither the employees nor their immediate supervisors were aware of departures from "business as usual." In the experiment, the employer varied the fraction of bad calls that were reported back to employees and supervisors at each call center. To more precisely estimate the true rate of fictitious pledges, the firm simultaneously increasing the true callback rate from 10% to 25% of pledges. By working closely with the firm, the researchers were able to collect primary survey data on employee attitudes toward the job, their expected job tenure, and the perceived difficulty of finding another, comparable job. These relate closely to the underlying theory of rational shirking behavior. This information is used to test whether those for whom the job was most valuable were also the employees least likely to engage in opportunistic behavior.

The main results are as follows. First, a significant fraction of employees behave according to the predictions of the rational cheater model. In particular, employees respond to a reduction in the perceived cost of opportunistic behavior by increasing the rate at which they shirk. Using the survey data collected, the authors find the employees who responded to reductions in monitoring tended to be those who perceived the employer as being unfair and uncaring. On the other hand, there is no evidence that individuals with good outside options increased shirking by more than other workers when the rate of monitoring declined. Second, a substantial proportion of employees do

[73] Olken (2007) presents evidence from a natural field experiment on the effects of top-down monitoring relative to grassroots participation on reducing corruption on road projects organized by village committees in Indonesia. Top-down monitoring via government audits is found to be the far more effective means of reducing corruption.

not appear to respond at all to manipulations in the monitoring rate. As with responses to monetary and non-monetary incentives documented above, there is considerable heterogeneity in how workers respond to employer monitoring. This underlying heterogeneity highlights the need to balance the need to reduce the shirking behavior of some workers inclined to rationally cheat, against those that are unlikely to do so under normal circumstances.

4.4. Moving forward

Economists have only recently begun to exploit field experiments in firms. This nascent literature has already highlighted the strengths of this methodology in being closely linked to testing alternative theories of individual behavior, of utilizing field experiments and structural modelling to make inference on the optimal design of incentive schemes, and to collect primary data to check for non-expected responses on other margins such as the quality of work, or to probe specific tests of the theory. We conclude by highlighting a few key areas for future work to consider.

First, the set of field experiments discussed have focused primarily, although not exclusively, on job tasks in which productivity is easy to observe, measure, compare across workers and time, and the quality of work performed is relatively easily monitored by management and assignable to individuals. Yet many jobs in the economy, or at higher tiers of firms' hierarchies, do not share such characteristics, and more research is required in such settings where performance is evaluated more subjectively, and might therefore be subject to influence activities (Milgrom, 1988), or favoritism (Prendergast and Topel, 1996). As primary data collection is part of the field experimenter's arsenal, this approach might especially help to shed light on these types of evaluations and incentive structures.

Second, most field experiments have been implemented to evaluate the effects of one time changes in management practices. Standard theory suggests history does not matter and that these effects should be equal and opposite to changing incentives the other way. This would be relatively straightforward to test, conditional on being able to control for natural time effects on behavior. A rejection of the standard model might then imply there can be persistent effects of short run changes in management practice. Such effects might operate through habit formation or reference point effects for example, that have been found using non-experimental data from real world settings (Mas, 2006).

Third, given the progression of field experiments exploring the effects of incentives on bottom-tier workers, and then to managers in the middle tier of the firm hierarchy, it is natural to ask whether field experiments might in the future extend to understanding executive pay. The last two decades have seen a surge in the popularity of performance pay for individuals in executive and managerial positions, from CEOs down to middle and lower management (Hall and Liebman, 1998; Hall and Murphy, 2003;

Oyer and Schaefer, 2004). However as yet there remains mostly an unwillingness of organizations to experiment in relation to such high stakes positions.[74]

Broader methodological issues remain to be borne in mind with regards to field experiments in firms. First, there are concerns over whether the set of firms and organizations that allow field experiments to be conducted within them, are selected in some way. For example, those firms that are most likely to gain from changes in management practices might be most amenable to field experiments on these dimensions. Given the potential for such non-random selection, field experiments ought to be designed to precisely measure differential effects, and less weight given to the levels effects.

Second, this body of field experiments offers an intriguing insight into whether firms choose their management practices optimally. Certainly, Shearer's (2004) study highlights why the firm was using piece rates and not fixed wages. For the firm studied in Bandiera et al., in each case the firm followed up on the results of the field experiment by maintaining the incentives that were introduced. However we have to be careful that while field experiments have focused on the effects of carefully engineered interventions on productivity, the firm chooses practices to maximize discounted profits. Productivity increases need to translate in profit increases. An example of this is in the study by Freeman and Kleiner (2005) on a US shoe manufacturer, who find that the move from piece rates to hourly wages reduced productivity, but increased the quality of work to such an extent that profits rose overall. Clearly, there remains scope for experimentation within firms to help them learn the optimal behaviors, and for this to have a large impact economy-wide, and perhaps go some way to explaining large productivity differences across otherwise observationally similar firms.

5. HOUSEHOLDS

Much of an individual's life cycle is spent in some form of partnership or family union. Despite widespread social changes in family structure in Western economies, families and multi-member households remain a key building block of society. Understanding how households make decisions has implications for many of the choices we have already touched upon, such as educational choices for children, labor market participation and labor supply. Shedding light on the household decision making processes also has profound implications for understanding whether, and how, policies such as income transfers and the regulation of marriage and divorce marriage markets, shape these outcomes.

The benchmark model of household behavior has been the unitary model, pioneered by Samuelson (1956) and Becker (1981). While this generates a rich set of predictions

[74] A similar set of issues arise for field experiments in public economics. In particular, understanding why individuals give to fundraisers or charitable causes. Large scale field experiments have so far focused on how to induce members of the public or those with affinity to the fundraising organization to give. However a disproportionate amount of funds raised come from a few very wealthy donors. No field experiments have been run on them.

for price and income effects on household behaviors, it remains silent on how conflicts between spouses are resolved. Modelling household decision making as the outcome of a bargaining process provides a natural way in which to introduce conflicts (Manser and Brown, 1980; McElroy and Horney, 1981; Chiappori, 1988). Hence, where these approaches differ is in whether households maximize according to a common or dictatorial set of preferences—the unitary approach—or whether they seek to maximize a weighted sum of household member preferences—the basis of the bargaining approach. On the other hand, a key feature of both modelling frameworks is that households are assumed to make efficient decisions.

Households might reasonably be expected to reach efficient outcomes because they have repeated and long term interactions, in strategic environments characterized by perfect information, and have the ability to communicate costlessly. Nevertheless, a more recent strand of the literature has developed that takes seriously the idea that either household members behave non-cooperatively within marriage (Ulph, 1988; Chen and Woolley, 2001), have private information or an inability to communicate perfectly (Pahl, 1983; Ligon, 1998; Goldstein and Udry, 1999; Boozer and Goldstein, 2003; Dubois and Ligon, 2004), or cannot make binding agreements (Lundberg and Pollak, 2003; Basu, 2006; Mazzocco, 2004, 2007; Rasul, 2008). In each case, household decisions can then be inefficient.

There are two long-standing strands of the empirical literature on household decision making that stem from these views of the world. First, there have been a number of attempts to uncover whether households bargain efficiently, as is implied by both the unitary and collective choice models. Many of these tests take the form of examining patterns of household demand and consumption (Browning and Chiappori, 1998) or testing for the equality of the marginal product of labor of household members across economic activities (Udry, 1994; Akresh, 2005). A first generation of field experiments on households has begun to shed light on this issue.[75]

Second, there is an older strand of the literature that uses non-experimental approaches to test for the assumption on whether households pool income, consistent with the predictions of the unitary framework, or whether the identity of the income earner matters for outcomes (Thomas, 1990, 1994; Hoddinott and Haddad, 1995; Duflo, 2003; Duflo and Udry, 2004; Rangel, 2006). Rather surprisingly given the roots of field experiments in the social experiments of the 1970s, relatively fewer field

[75] Tests based on demand patterns exploit the fact that utility maximization by a single consumer subject to a linear budget constraint implies Slutsky symmetry, namely the restriction of symmetry on the matrix of compensated price responses. This prediction is typically rejected in household data (Deaton, 1990; Browning and Meghir, 1991; Banks et al., 1997; Browning and Chiappori, 1998). Browning and Chiappori (1998) derive the counterpart to the Slutsky matrix for multi-member households solely under the assumption of efficient within-household decision making, consistent with Nash bargaining models. They show the assumption of efficiency generates testable restrictions on household demand functions, and distinguish the collective model from both the unitary and the entirely unrestricted case.

experiments have been conducted to help test the specific predictions of either unitary or collective bargaining frameworks.

A parallel stream of literature relates to the use of social experiments to evaluate conditional cash transfer programs, which was touched upon earlier. Two notable studies that have used data from the *PROGRESA* intervention in rural Mexico are Attanasio et al. (2006) and Todd and Wolpin (2006). These both combine the experimental variation in *PROGRESA* transfers across randomly assigned villages with structural estimation of a household's dynamic behavior to shed light on outcomes under alternative policy designs.

5.1. Efficiency

Ashraf (2009) presents evidence from a framed field experiment to understand how information and communication affect household financial decisions. The experiment was conducted with a sample of current or former clients of a rural bank in the Philippines. The main decision each subject had to make was over whether to spend or save income received during the experiment. More precisely, subjects had to choose how to allocate 200 pesos received between: (i) direct deposits into their own or a joint account; (ii) committed consumption using redeemable gift certificates. Each subject was randomly assigned, with his or her spouse, to one of three treatments that varied the privacy of information spouses had, and the ability of spouses to communicate with each other. 149 married couples are involved in the experiment.[76]

In the first treatment, subjects are separated from their spouses at the outset of the experiment. This treatment is referred to as "private information without pre-play communication". Under this treatment spouses have no information on whether and how much income is received by the spouse, what decisions they have made, or the outcomes obtained. In the second treatment spouses learn each others' payoffs and choice sets. In this treatment, referred to as "public information without pre-play communication", spouses make simultaneous decisions and so cannot communicate nor observe each others decisions *ex ante*. In the final treatment the procedure is as in the previous treatment except that spouses are able to communicate before making their decisions, and their decisions are observable to each other. This is referred to as the "public treatment".

Clearly, in the absence of a field experiment, trying to uncover and exploit plausibly exogenous sources of variation in the information available to spouses or their ability to communicate is difficult, and likely to be correlated to factors that affect outcomes

[76] As in any framed field experiment or laboratory experiment, subjects need to be recruited. Framed field experiments that aim to replicate natural settings—say by working in conjunction with local organizations—might provide data from which to assess whether participants differ from those that choose not to participate. As discussed earlier, the nature of self-selection into experiments is a phenomenon that is only beginning to be understood (Lazear et al., 2009). Equally important, given the relatively small sample sizes inherent to many framed field experiments, it is crucial to be clear on how large a sample would be required to detect statistically different observable characteristics between participants and non-participants.

directly. Hence this research design allows economists to more carefully scrutinize causal changes in behavior along dimensions that are theoretically important, yet empirically almost impossible to measure in the absence of a field experiment. However, as with the other settings considered throughout, field experiments conducted with households raise important issues that need to be taken into consideration when interpreting results.

First, in common with laboratory experiments, behavior in framed field experiments might not mimic behavior in the real world. Ashraf (2009) addresses this concern by running the experiment in conjunction with a rural bank that all participants were familiar with, and by designing treatments that capture real world differences in communication and information across households in this setting. Second, households are engaged in repeated interactions outside of the context of the field experiment. Hence behavior within an experiment can be undone, or potentially reinforced, by behavior outside of the experiment. To try and address this issue, Ashraf provides payoffs in the form of person-specific gift certificates. Both methodological issues need to be considered in all field experiments with households.

Ashraf's (2009) results shed light on the interplay between information, communication and gender in household decision making. Relative to field experiments in other settings, when the experimenter is engaged in primary data collection with households it is of even greater importance to understand societal norms of behavior within marriage. For example, in the Philippines, women are typically in charge of the financial management of the household, making key decisions on budgeting and allocation. Understanding the context in which the experiment takes place is crucial for designing treatments that reflect real world trade-offs subjects face, and to closely align experimental designs with a theoretical framework. Of course, the cost of this precision in any given context is the limited ability, all else equal, to extrapolate findings to households operating under very different norms.

The three main results are as follows. First, men are found to be more likely to deposit money into their own account under the private treatments, and are more likely to commit it to consumption under the public treatment. Second, the differences in behavior by gender are subtle. A subset of women—those whose husbands normally control household savings decisions—behave in the same way as men whose wives normally control household savings decisions. Third, communication between spouses at the time of decision making induces the majority of men to place the income into their spouses account rather than consume it or put it into their own account. To understand these results, Ashraf discusses a framework of income monitoring within the household where observability of income and communication at the time financial decision making, significantly change the monitor's ability to enforce contracts. The results can then be understood as spouses responding strategically to changes in information and communication and contract enforceability. They suggests a specific channel through

which asymmetric information can create inefficient outcomes in financial decision making, by providing incentives to hide one's additional income from one's spouse.[77]

5.2. Moving forward

Labor economists have sought to explain a far richer set of research questions than just those related to behavior within households. Foremost among these other issues has been the research into the causes and consequences of the formation and dissolution of households. Field experiments have recently begun to help explore issues related to the formation of households or partnerships in the first place. Two examples are Fisman et al. (2006) and Fisman et al. (2008) who conduct a framed field experiment to measure differential preferences in dating across genders and races, respectively. To do so, both studies analyze individual choices of subjects in an experimental speed dating game.[78]

On differential preferences across genders, Fisman et al. (2006) find that women place greater weight on the intelligence and the race of partner, while men respond more to physical attractiveness. Moreover, men do not value women's intelligence or ambition when it exceeds their own.

On racial preferences, Fisman et al. (2008) find that there is a strong asymmetry in racial preferences across genders: women of all races exhibit strong same race preferences, while men—of all races —do not. Second, subjects' background influences their racial preferences: subjects that come from locations that are measured to be more racially intolerant, using data from the General Social Survey and World Values Surveys, reveal stronger preferences for same race preferences. This is despite the subject pool being drawn from individuals that currently reside away from home, and attend a top US university. Third, those exposed to other races in early life—as measured by the fraction of individuals of a given race in the zip code where the subject grew up—are less willing to date someone from this race, suggesting that familiarity might reduce racial tolerance. Finally, physically more attractive individuals are less sensitive to the race of potential partners in the experiment.

This experimental approach provides a nice complement to other non-experimental studies applying structural methods to estimate similar preference parameters in the context of online dating services (Hitsch et al., 2010). Given the growth in availability of online data in economic research, perhaps in the near future we will witness research methods combining field experiments with interventions akin to audit studies that were previously discussed in relation to the economics of discrimination.

77 This strand of field experiments is growing for example, Robinson (2008) presents evidence from a framed field experiment on 142 households in Kenya to test whether intra-household risk sharing arrangements are efficient, and if not, whether limited commitment caused by contractual incompleteness partially explains behavior.

78 Stevenson and Wolfers (2007) provide a recent overview of the most pressing issues that are being addressed in research in the economics of the family.

As yet though, on many aspects of the formation and dissolution of families, few research designs have credibly exploited experimental sources of variation from which to identify causal effects. The nature of questions involved might mean these sets of research questions remain outside the domain of field experiments.

6. CONCLUDING REMARKS

Given that complexities of markets severely constrain the ability of traditional economic tools to examine behavioral relationships, it is not surprising that economists have increasingly turned to experimental methods. Within this recent trend is a relatively new approach—field experiments—which have dramatically risen in popularity over the past several years. Since field experiments will likely continue to grow in popularity as scholars continue to take advantage of the settings where economic phenomena present themselves, we view this study as an opportunity to step back and discuss a few of the areas within labor economics wherein field experiments have contributed to our economic understanding. Our central task is to highlight what we view to be the central advantages of the field experimental approach: (i) using economic theory to design the null and alternative hypotheses; (ii) engineering exogenous variation in real world economic environments to establish causal relations and learn the mechanisms behind them; and (iii) engaging in primary data collection and often working closely with practitioners.

A second goal of this study is to draw attention to a methodological contribution of field experiments: complementing other empirical approaches and allowing an exploration of the generalizability of behaviors across settings, such as lab and field behavior. When taking account of the stock of evidence, it becomes clear how field experiments can play an important role in the discovery process by allowing one to make stronger inference than can be achieved from lab or uncontrolled data alone. In this way, the various empirical approaches should be thought of as strong complements—much like theory and empirical modeling—and combining insights from each of the methodologies will permit economists to develop a deeper understanding of our science.

REFERENCES

Abdulkadiroglu, A., Angrist, J.D., Dynarski, S.M., Kane, T.J., Pathak, P., 2009. Accountability and flexibility in public schools: evidence from Boston's charters and pilots. National Bureau of Economic Research Working Paper No. 15549.

Acemoglu, D., 2002. Technical change, inequality, and the labor market. Journal of Economic Literature 40, 7–72.

Akerlof, G.A., 1980. The theory of social custom, of which unemployment may be one consequence. Quarterly Journal of Economics 94, 749–775.

Akerlof, G.A., 1982. Labor contracts as a partial gift exchange. Quarterly Journal of Economics 97, 543–569.

Akerlof, G.A., Yellen, J.L., 1988. Fairness and unemployment. American Economic Review Papers and Proceedings 78, 44–49.

Akerlof, G.A., Yellen, J.L., 1990. The fair wage-effort hypothesis and unemployment. Quarterly Journal of Economics 105, 255–283.

Akerlof, G.A., Kranton, R., 2005. Identity and the economics of organizations. Journal of Economic Perspectives 19, 9–32.
Akresh, R., 2005. Understanding Pareto inefficient intrahousehold allocations. Institute for the Study of Labor Discussion Paper 1858.
Alabarran, P., Attanasio, O., 2003. Limited commitment and crowding out of private transfers: evidence from a randomised experiment. Economic Journal 113, C77–C85.
Alchian, A.A., Demsetz, H., 1972. Production, information costs, and economic organizations. American Economic Review 62, 777–795.
Alevy, J., Haigh, M., List, J.A., 2007. Information cascades: evidence from a field experiment with financial market professionals. Journal of Finance 62, 151–180.
Altonji, J.G., Blank, R., 1999. Race and gender in the labor market. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 3C. Elsevier Science B.V., pp. 3143–3259.
Altonji, J.G., Pierret, C., 2001. Employer learning and statistical discrimination. Quarterly Journal of Economics CXVI, 293–312.
Anderson, M.L., 2008. Multiple inference and gender differences in the effects of early intervention: a reevaluation of the Abecedarian, Perry preschool, and Early Training projects. Journal of the American Statistical Association 103, 1481–1495.
Andreoni, J., Petrie, R., 2004. Public goods experiments without confidentiality: a glimpse into fund-raising. Journal of Public Economics 88, 1605–1623.
Angrist, J.D., Krueger, A.B., 1999. Empirical Strategies in Labor Economics. In: Handbook of Labor Economics, vol. 3, Part A. Elsevier, pp. 1277–1366 (Chapter 23).
Angrist, J.D., Krueger, A.B., 2001. Instrumental variables and the search for identification: from supply and demand to natural experiments. Journal of Economic Perspectives 15, 69–85.
Angrist, J.D., Lang, K., 2004. Does school integration generate peer effects? Evidence from Boston's Metco program. American Economic Review 94 (5), 1613–1634.
Angrist, J.D., Lavy, V., 1999. Using Maimonides' rule to estimate the effect of class size on scholastic achievement. Quarterly Journal of Economics 114, 533–575.
Angrist, J.D., Lavy, V., 2002. New evidence on classroom computers and pupil learning. Economic Journal 112, 735–765.
Angrist, J.D., Lavy, V., 2009. The effects of high stakes high school achievement awards: evidence from a randomized trial. American Economic Review 99 (4), 1384–1414.
Angrist, J.D., Bettinger, E., Bloom, E., King, E., Kremer, M., 2002. Vouchers for private schooling in colombia: evidence from a randomized natural experiment. American Economic Review 92, 1535–1558.
Angrist, J.D., Bettinger, E., Kremer, M., 2006. Long-term educational consequences of secondary school vouchers: evidence from administrative records in Colombia. American Economic Review 96, 847–862.
Angrist, J.D., Barski, S.M., Kane, T.J., Pathak, P., Walters, C.R., 2010. Who benefits from KIPP? National Bureau of Economic Research Working Paper No. 15740.
Aoyagi, M., 2007. Information feedback in a dynamic tournament. mimeo, Osaka University.
Arabsheibani, G.R., Marin, A., Wadsworth, J., 2005. Gay pay in the UK. Economica 72, 333–347.
Arcidiacono, P., 2005. Affirmative action in higher education: how do admission and financial aid rules affect future earnings? Econometrica 73, 1477–1524.
Arrow, K., 1972. The theory of discrimination. In: Ashenfelter, O., Rees, A. (Eds.), Discrimination in Labor Markets. Princeton University Press, Princeton, NJ.
Ashenfelter, O., 1990. Nonparametric estimates of the labor-supply effects of negative income tax programs. Journal of Labor Economics 8, S396–S415.
Ashraf, N., 2009. Spousal control and intra-household decision making: an experimental study in the philippines. American Economic Review 99, 1245–1277.
Attanasio, O., Barr, A., Cardenas, J.C., Genicot, G., Meghir, C., 2009. Risk pooling, risk preferences, and social networks. mimeo, Oxford University.
Attanasio, O., Meghir, C., Santiago, A., 2006. Education choices in Mexico: using a structural model and a randomised experiment to evaluate *progresa*. IFS Working Paper.
Autor, D., Katz, L., Kearney, M., 2005. Rising wage inequality: the role of composition and prices. mimeo, Harvard University.

Azmat, G., Iriberri, N., 2009. The importance of relative performance feedback information: evidence from a natural experiment using high school students. mimeo, Universitat Pompeu Fabra.
Baker, G., Jensen, M.C., Murphy, K.J., 1988. Compensation and incentives: practice vs. theory. Journal of Finance 43, 593–616.
Ball, S., Eckel, C., Grossman, P.J., Zame, W., 2001. Status in markets. Quarterly Journal of Economics 116, 161–188.
Bandiera, O., Barankay, I., Rasul, I., 2005. Social preferences and the response to incentives: evidence from personnel data. Quarterly Journal of Economics 120, 917–962.
Bandiera, O., Barankay, I., Rasul, I., 2007. Incentives for managers and inequality among workers: evidence from a firm level experiment. Quarterly Journal of Economics 122, 729–774.
Bandiera, O., Barankay, I., Rasul, I., 2009a. Team incentives: evidence from a field experiment, mimeo, University College London.
Bandiera, O., Barankay, I., Rasul, I., 2010. Social incentives in the workplace. Review of Economic Studies 77 (April), 417–459.
Bandiera, O., Larcinese, V., Rasul, I., 2009b. Blissful ignorance? Evidence from a natural experiment on the effect of individual feedback on performance. mimeo, LSE.
Bandiera, O., Larcinese, V., Rasul, I., 2010. Heterogeneous class size effects: new evidence from a panel of university students. Economic Journal (forthcoming).
Banerjee, A.V., Suraj, J., Kremer, M., Lanjouw, J., Lanjouw, P., 2001. Promoting school participation in rural rajasthan: results from some prosepctve trials. mimeo, Massachusetts Institute of Technology.
Banerjee, A.V., Cole, S., Duflo, E., Linden, L., 2007. Remedying education: evidence from two randomized experiments in India. Quarterly Journal of Economics 122, 1235–1264.
Banks, J., Blundell, R., Lewbel, A., 1997. Quadratic Engel curves and consumer demand. Review of Economics and Statistics 79, 527–539.
Barnard, C., 1938. The Functions of the Executive. Harvard University Press, Cambridge.
Baron, J.N., Kreps, D.M., 1999. Strategic Human Resources: Frameworks for General Managers. John Wiley and Sons, Inc., New York.
Baron, J.N., Pfeffer, J., 1994. The social psychology of organizations and inequality. Social Psychology Quarterly 57, 190–209.
Barr, A., Serneels, P., 2009. Reciprocity in the workplace. Experimental Economics 12, 99–112.
Barrera-Osorio, F., Bertrand, M., Linden, L., Perez-Calle, F., 2008. Conditional cash transfers in education: design features. Peer and Sibling Effects: Evidence from a Randomized Experiment in Colombia, The World Bank: Impact Evaluation Series No. 20.
Basu, K., 2006. Gender and say: a model of household behavior with endogenously-determined balance of power. Economic Journal 116, 558–580.
Becker, G.S., 1957. The Economics of Discrimination, second ed., University of Chicago Press, Chicago.
Becker, G.S., 1981. A Treatise on the Family. Harvard University Press, Cambridge.
Beer, M., 1990. Performance Appraisal. mimeo, Harvard Business School.
Behrman, J.R., Cheng, Y., Todd, P.E., 2004. Evaluating preschool programs when length of exposure to the program varies: a nonparametric approach. Review of Economics and Statistics 86, 108–132.
Benabou, R., Tirole, J., 2000. Self-confidence and social interactions. NBER Working Paper 7585.
Benabou, R., Tirole, J., 2003. Intrinsic and extrinsic motivation. Review of Economic Studies 70, 489–520.
Berg, N., Lien, D., 2002. Measuring the effect of sexual orientation on income: evidence of discrimination? Contemporary Economic Policy 20, 394–414.
Bernheim, D., 1994. A theory of conformity. Journal of Political Economy 102, 841–877.
Berry, J., 2009. Child control in education decisions: an evaluation of targeted incentives to learn in India. mimeo, Massachusetts Institute of Technology.
Bertrand, M., Mullainathan, S., 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. American Economic Review 94, 991–1013.
Bertrand, M., Schoar, A., 2003. Managing with style: the effect of managers on firm policies. Quarterly Journal of Economics 118, 1169–1208.
Besley, T.J., Ghatak, M., 2005. Competition and incentives with motivated agents. American Economic Review 95, 616–636.

Besley, T.J., Ghatak, M., 2008. Status incentives. American Economic Review Papers and Proceedings 98, 206–211.
Bewley, T.F., 1999. Why Wages Don't Fall During a Recession. Harvard University Press, Cambridge.
Bhaskar, V., 2002. Relative performance evaluation and limited liability. mimeo, University of Essex.
Bhattacharya, D., Dupas, P., 2010. Inferring welfare maximizing treatment assignment under budget constraints. NBER Working Paper 14447.
Bjorklund, A., Regner, H., 1996. Experimental evaluation of european labour market policy. In: Scmid, G., O'Reilly, J. (Eds.), International Handbook of Labour Market Policy and Evaluation. pp. 89–115 (Chapter 3).
Blanes-i-Vidal, J., Nossol, M., 2009. Tournaments without prizes: evidence from personnel records. mimeo LSE.
Bloom, N., Van Reenen, J., 2011. Human resource management and productivity. In: New Developments and Research on Labor Markets, first ed., In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 4B. Elsevier, pp. 1697–1767 (Chapter 19).
Blundell, R., Costa-Dias, M., 2002. Alternative approaches to evaluation in empirical microeconomics. Portuguese Economic Journal 1, 91–115.
Bleakley, H., 2007. Disease and development: evidence from hookworm eradication in the American South. Quarterly Journal of Economics 122, 73–112.
Bobonis, G.J., Miguel, E., Puri-Sharma, C., 2006. Anemia and school participation. Journal of Human Resources 41, 692–721.
Bognanno, M.L., 2001. Corporate tournaments. Journal of Labor Economics 19, 290–315.
Bohm, P., 1984. Revealing demand for an actual public good. Journal of Public Economics 24, 135–151.
Boozer, M., Goldstein, M.P., 2003. Poverty measurement and dynamics. mimeo, Yale University.
Bramel, D., Friend, R., 1981. Hawthorne, the myth of the docile worker, and class bias in psychology. American Psychologist 36, 867–878.
Brown-Kruse, J., Cronshaw, M.B., Schenk, D.J., 2007. Theory and experiments on spatial competition. Economic Inquiry 31, 139–165.
Browning, M., Chiappori, P.-A., 1998. Efficient intra-household allocations: a general characterisation and empirical tests. Econometrica 66, 1241–1278.
Browning, M., Meghir, C., 1991. The effects of male and female labor supply on commodity demands. Econometrica 59, 925–951.
Bull, C., Schotter, A., Weigelt, K., 1987. Tournaments and piece rates: an experimental study. Journal of Political Economy 95, 1–33.
Burgess, S., Propper, C., Slater, H., Wilson, D., 2005. Who wins and who loses from school accountability? The distribution of educational gain in English secondary schools. CMPO Working Paper Series 05/128.
Butler, R., 1987. Task-involving and ego-involving properties of evaluation: effects of different evaluation conditions on motivational perceptions, interest, and performance. Journal of Educational Psychology 79, 474–482.
Cadsby, C.B., Song, F., Tapon, F., 2007. Sorting and incentive effects of pay for performance: an experimental investigation. Academy of Management Journal 50, 387–405.
Cahuc, P., Postel-Vinay, F., Robin, J.-M., 2006. Wage bargaining with on-the-job search: a structural econometric model. Econometrica 74, 323–364.
Cameron, J., Pierce, W.D., 1994. Reinforcement, reward, and intrinsic motivation: a meta-analysis. Review of Educational Research 64, 363–423.
Cameron, S.V., Heckman, J.J., 2001. The dynamics of educational attainment for black, hispanic, and white males. Journal of Political Economy 109, 455–499.
Cameron, S.V., Taber, C., 2004. Estimation of educational borrowing constraints using returns to schooling. Journal of Political Economy 112, 132–182.
Card, D., 1999. The causal effect of education on earnings. In: Handbook of Labor Economics, first ed., In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 3. Elsevier, pp. 1801–1863 (Chapter 30).
Card, D., Krueger, A.B., 1992. School quality and black-white relative earnings: a direct assessment. Quarterly Journal of Economics 107, 151–200.

Carpenter, J., Seki, E., 2010. Do social preferences increase productivity? Field experimental evidence from fisherman in Toyama bay. Economic Inquiry, doi:10.1111/j.1465-7295.2009.00268.x.
Case, A., Deaton, A., 1999. School inputs and educational outcome in South Africa. Quarterly Journal of Economics 114, 1047–1084.
Charles, K.K., Guryan, J., 2008. Prejudice and Wages: an empirical assessment of Becker's The Economics of Discrimination. Journal of Political Economy 116, 773–809.
Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. Quarterly Journal of Economics 117, 817–869.
Charness, G., Yang, C.-L., 2008. Endogenous group formation and public goods provision: exclusion, exit, mergers, and redemption. UC Santa Barbara Department of Economics Working Paper 13-08.
Charness, G., Kuhn, P., 2010. Lab labor: what can labor economists learn from the lab? NBER Working Paper 15913.
Charness, G., Kuhn, P., 2011. Lab labor: What can labor economists learn from the lab? In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 4A. Elsevier, pp. 229–331.
Charness, G., Villeval, M.-C., 2009. Cooperation and competition in intergenerational experiments in the field and laboratory. American Economic Review 99, 956–978.
Chen, Z., Woolley, F., 2001. A Cournot-Nash model of family decision making. Economic Journal 111, 722–748.
Chiappori, P.-A., 1988. Rational household labor supply. Econometrica 56, 63–90.
Chiappori, P.-A., Salanie, B., 2003. Testing contract theory: a survey of some recent work. In: Dewatripont, M., Hansen, L.P., Turnovsky, S.J. (Eds.), Advances in Economics and Econometrics Theory and Applications: Eighth World Congress Volume 1. Cambridge University Press, Cambridge.
Cipriani, M., Guarino, A., 2009. Herd behavior in financial markets: an experiment with financial market professionals. Journal of the European Economic Association 7, 206–233.
Clark, D., 2009. The performance and competitive effects of school autonomy. Journal of Political Economy 117, 745–783.
Coate, S., Loury, G.C., 1993. Will affirmative-action policies eliminate negative stereotypes? American Economic Review 83, 1220–1240.
Coleman, J.S., et al., 1966. Equality of Educational Opportunity. US Government Printing Office, Washington, DC.
Cornwell, C., Mustard, D.B., Sridhar, D.J., 2006. The enrollment effects of merit-based financial aid: evidence from Georgia's HOPE program. Journal of Labor Economics 24, 761–786.
Cullen, J.B., Jacob, B.A., Levitt, S.D., 2006. The effect of school choice on participants: evidence from randomized lotteries. Econometrica 74, 1191–1230.
Cunha, F., Heckman, J.J., 2009. The economics of psychology of inequality and human development. Journal of the European Economic Association 7, 320–364.
Cooper, D.J., Kagel, J.H., Lo, W., Gu, Q.L., 1999. Gaming against managers in incentive systems: experimental results with Chinese students and Chinese managers. American Economic Review 89, 781–804.
Currie, J., 2001. Early childhood education programs. Journal of Economic Perspectives 15, 213–238.
Currie, J., Thomas, D., 1995. Does head start make a difference? American Economic Review 85 (3), 341–364.
Currie, J., Thomas, D., 2000. School quality and the longer term effects of head start. Journal of Human Resources 35, 755–774.
Deaton, A.S., 1990. Price elasticities from survey data: extensions and Indonesian results. Journal of Econometrics 44, 281–309.
Dobbie, W., Fryer, R.G., 2009. Are high quality schools enough to close the achievement gap? Evidence from social experiment in Harlem. National Bureau of Economic Research Working Paper No. 15473.
De Giorgi, G., Pellizzari, M., Woolston, W.G., 2009. Class-size and class heterogeneity. mimeo, Stanford University.
Deci, E.L., Koestner, R., Ryan, R.M., 2001. Extrinsic rewards and intrinsic motivation in education: reconsidered once again. Review of Educational Research 71, 1–27.
Delfgaauw, J., Dur, R., 2004. Incentives and workers' motivation in the public sector. Tinbergen Institute Discussion Papers 04-060/1.

Dewatripont, M., Jewitt, I., Tirole, J., 1999. The economics of career concerns, Part II: application to missions and accountability of government agencies. Review of Economic Studies 66, 199–217.
Ding, M., Grewal, R., Liechty, J., 2005. Incentive-aligned conjoint analysis. Journal of Marketing Research 42, 67–83.
Dingwall, R., 1980. Ethics and Ethnography. Sociological Review 28, 871–891.
Dixit, A., 2002. Incentives and organizations in the public sector: an interpretative review. Journal of Human Resources 37, 696–727.
Dohmen, T.J., Falk, A., 2006. Performance pay and multi-dimensional sorting: productivity, preferences and gender. Institute for the Study of Labor Discussion Paper No. 2001.
Doolittle, F., Traeger, L., 1990. Implementing the national JPTA study. Department of Labor, Washington DC.
Dubois, P., Ligon, E., 2004. Incentives and nutrition for rotten kids: intrahousehold food allocation in the Phillipines. mimeo, UC Berkeley.
Duflo, E., 2003. Grandmothers and granddaughters: old age pension and intra-household allocation in South Africa. World Bank Economic Review 17, 1–25.
Duflo, E., Udry, C., 2004. Intrahousehold resource allocation in Cote d'Ivoire: social norms. Separate Accounts, and Consumption Choices, NBER Working Paper No. 10498.
Duflo, E., Gale, W., Iebman, J., Orszag, P., Saez, E., 2006. Saving incentives for low- and middle-income families: evidence from a field experiment with H&R block. Quarterly Journal of Economics 121, 1311–1346.
Duflo, E., Dupas, P., Kremer, M., 2009. Peer effects, teacher incentives and the impact of tracking: evidence from a randomized evaluation in Kenya. National Bureau of Economic Research Working Paper No. 14475.
Dunford, F.W., 1990. Random assignment: practical considerations from field experiments. Evaluation and Program Planning 13, 125–132.
Duesenberry, J.S., 1949. Income, Saving, and the Theory of Consumer Behavior. Harvard University Press, Cambridge, MA.
Ederer, F., 2008. Feedback and Motivation in Dynamic Tournaments, mimeo, MIT.
Ederer, F., Fehr, E., 2007. Deception and incentives: how dishonesty undermines effort provision. IZA Discussion Paper 3200.
Ellingsen, T., Johannesson, M., 2008. Pride and prejudice: the human side of incentive theory. American Economic Review 98, 990–1008.
Encinosa, W.E., Gaynor, M.S., Rebitzer, J.B., 1997. The sociology of groups and the economics of incentives: theory and evidence on compensation systems. National Bureau of Economic Research Working Paper 5953.
Engle, P., Black, M., Behrman, J., Cabral De Mello, M., Gertler, P., Kapiriri, L., Martorell, R., Young, M., 2007. Strategies to avoid the loss of developmental potential in more than 200 million children in the developing world. The Lancet 369, 229–242.
Epple, D., Romano, R.E., 1998. Competition between private and public schools, vouchers, and peer-group effects. American Economic Review 88, 33–62.
Epple, D., Romano, R.E., Sieg, H., 2006. Admission, tuition, and financial aid policies in the market for higher education. Econometrica 74, 885–928.
Ehrenberg, R., Bognanno, M., 1990. The incentive effects of tournaments revisited: evidence from the european PGA tour. Industrial Labor Relations Review 43, 74–89.
Eriksson, T., 1999. Executive compensation and tournament theory: empirical tests on Danish data. Journal of Labor Economics 17, 262–280.
Eriksson, T., Poulsen, A., Villeval, M.-C., 2008a. Feedback and incentives: experimental evidence. IZA Discussion Paper 3440.
Eriksson, T., Teyssier, S., Villeval, M.-C., 2008b. Self-selection and the efficiency of tournaments. Economic Inquiry 47, 530–548.
Ferber, R., Hirsch, W.Z., 1982. Social Experimentation and Economic Policy. Cambridge University Press, London.
Fehr, E., Falk, A., 1999. Wage rigidity in a competitive incomplete contract market. Journal of Political Economy 107, 106–134.

Fehr, E., Fischbacher, U., 2002. Why social preferences matter – the impact of non–selfish motives on competition, cooperation and incentives. Economic Journal 112, C1–C33.
Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition and cooperation. Quarterly Journal of Economics 114, 817–868.
Fehr, E., Schmidt, K.M., 2000. Fairness, incentives, and contractual choices. European Economic Review 44 (4–6), 1057–1068.
Fehr, E., Kirchsteiger, G., Riedl, A., 1993. Does fairness prevent market clearing? An experimental investigation. Quarterly Journal of Economics 108, 437–459.
Fehr, E., Goette, L., Huffman, D., 2004. Loss aversion and labor supply. Journal of the European Economic Association 2, 216–228.
Ferreyra, M.M., 2007. Estimating the effects of private school vouchers in multidistrict economies. American Economic Review 97, 789–817.
Fershtman, C., Hvide, H.K., Weiss, Y., 2003. Cultural Diversity, Status Concerns and the Organization of Work. CEPR Discussion Paper No. 3982.
Figlio, D., Getzler, L., 2002. Accountability, Ability and Disability: Gaming the System. National Bureau of Economic Research Working Paper 9307.
Figlio, D.N., Kenny, L.W., 2007. Individual teacher incentives and student performance. Journal of Public Economics 91, 901–914.
Figlio, D.N., Rouse, C.E., 2006. Do accountability and voucher threats improve low-performing schools? Journal of Public Economics 90 (1–2), 239–255.
Fisher, R.A., 1926. The arrangement of field trials. Journal of the Ministry of Agriculture of Great Britain 33, 503–513.
Fisman, R., Iyengar, S., Kamenica, E., Simonson, I., 2006. Gender differences in mate selection: evidence from a speed dating experiment. Quarterly Journal of Economics 121, 673–697.
Fisman, R., Iyengar, S., Kamenica, E., 2008. Racial preferences in dating: evidence from a speed dating experiment. Review of Economic Studies 75, 117–132.
Francois, P., 2000. Public service motivation as an argument for government provision. Journal of Public Economics 78, 275–299.
Frank, R.H., 1985. Choosing the right pond: human behavior and the quest for status. Oxford University Press, New York.
Franke, R.H., Kaul, J.D., 1978. The Hawthorne experiments: first statistical interpretation. American Sociological Review 43, 623–643.
Freeman, R.B., 1987. Labour Economics. The New Palgrave Dictionary of Economics, first edition.
Freeman, R.B., Kleiner, M.M., 2005. The last American shoe manufacturers: decreasing productivity and increasing profits in the shift from piece rates to continuous flow production. Industrial Relations 44, 307–330.
Freeman, R.B., Gelber, A.M., 2008. Prize Structure and Information in Tournaments: Experimental Evidence. mimeo, Harvard University.
Friedman, M., Savage, L.J., 1948. The utility analysis of choice involving risk. Journal of Political Economy 56, 279–304.
Fryer, R.G., 2011. Racial inequality in the 21st century: the declining significance of discrimination. In: New Developments and Research on Labor Markets, first ed., In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 4B. Elsevier, pp. 855–971 (Chapter 10).
Gale, E.A.M., 2004. The Hawthorne studies—a fable for our times? Quarterly Journal of Medicine 97, 439–449.
Garces, E., Thomas, D., Currie, J., 2002. Longer-term effects of Head Start. American Economic Review 92, 999–1012.
Gibbons, R.S., 1987. Piece-rate incentive schemes. Journal of Labor Economics 5, 413–429.
Gibbons, R.S., Murphy, K.J., 1990. Relative performance evaluation for chief executive officers. Industrial Labor Relations Review 43, 30–52.
Gibbs, M., 1991. An Economic Approach to Process in Pay and Performance Appraisals. mimeo, University of Chicago GSB.
Gilbert, D.T., Pinel, E.C., Wilson, T.D., Blumberg, S.J., Wheatley, T.P., 1998. Immune neglect: a source of durability bias in affective forecasting. Journal of Personality and Social Psychology 75, 617–638.

Gillespie, R., 1991. Manufacturing Knowledge: A History of the Hawthorne Experiments. Cambridge University Press, New York.
Gine, X., Karlan, D., Zinman, J., 2007. The Risk of Asking: Measurement Effects from a Baseline Survey in an Insurance Takeup Experiment. Working paper, World Bank.
Glewwe, P., Kremer, M., 2006. Schools, teachers, and education outcomes in developing countries. In: Handbook of the Economics of Education. Elsevier, pp. 945–1017 (Chapter 16).
Glewwe, P., Kremer, M., Ilias, N., 2003. Teacher Incentives. National Bureau of Economic Research Working Paper No. 9671.
Glewwe, P., Kremer, M., Moulin, S., Zitzewitz, E., 2004. Retrospective vs. prospective analyses of school inputs: the case of flip charts in Kenya. Journal of Development Economics 74, 251–268.
Gneezy, U., List, J.A., 2006. Putting behavioral economics to work: testing for gift exchange in labor markets using field experiments. Econometrica 74, 1365–1384.
Gneezy, U., List, J.A., Price, M.K., 2010. Clean Evidence of Statistical Disrimination. Working Paper, University of Chicago.
Gneezy, U., Rustichini, A., 2000. Pay enough or don't pay at all. Quarterly Journal of Economics 115, 791–810.
Goldin, C., Rouse, C., 2000. Orchestrating impartiality: the impact of "blind" auditions on female musicians. American Economic Review 99, 715–741.
Goldstein, M.P., Udry, C.R., 1999. Gender and Land Resource Management in Southern Ghana. mimeo, Yale University.
Grantham-McGregor, S., Powell, C., Walker, S., Himes, J., 1991. Nutritional supplementation, psychosocial stimulation, and mental development of stunted children: the Jamaican study. The Lancet 338, 1–5.
Grantham-McGregor, S., Cheung, Y., Cueto, S., Glewwe, P., Richter, L., Strupp, B., 2007. Developmental potential in the first 5 years for children in developing countries. The Lancet 369, 60–70.
Green, J.R., Stokey, N.L., 1983. A comparison of tournaments and contracts. Journal of Political Economy 91, 349–364.
Greenberg, J., 1988. Equity and workplace status: a field experiment. Journal of Applied Psychology 73, 606–613.
Greenberg, D., Shroder, M., Onstott, M., 1999. The social experiment market. Journal of Economic Perspectives 13, 157–172.
Greenberg, D., Shroder, M., 2004. The Digest of Social Experiments. The Urban Institute Press, Washington.
Groves, T., Hong, Y., Mcmillan, J., Naughton, B., 1994. Autonomy and incentives in Chinese state enterprises. Quarterly Journal of Economics 109, 183–209.
Hall, B., Liebman, J., 1998. Are CEOs really paid like bureaucrats? Quarterly Journal of Economics 113, 653–691.
Hall, B., Murphy, K.J., 2003. The trouble with stock options. Journal of Economic Perspectives 17, 49–70.
Hamilton, B.H, Nickerson, J.A., Owan, H., 2003. Team incentives and worker heterogeneity: an empirical analysis of the impact of teams on productivity and participation. Journal of Political Economy 111, 465–497.
Hanushek, E.A., 1986. The economics of schooling: production and efficiency in public schools. Journal of Economic Literature 24, 1141–1177.
Hanushek, E.A., 2006. School resources. In: Handbook of the Economics of Education. Elsevier, pp. 865–908 (Chapter 14).
Hanushek, E.A., 2007. Some US evidence on how the distribution of educational outcomes can be changed. In: Schools and the Equal Opportunity Problem. MIT Press, pp. 159–190 (Chapter 7).
Hanushek, E.A., Raymond, M.E., 2004. The effect of school accountability systems on the level and distribution of student achievement. Journal of the European Economic Association 2, 406–445.
Harrison, G., List, J., 2004. Field experiments. Journal of Economic Literature 152, 1009–1055.
Hart, O., Holmstrom, B., 1987. The theory of contracts. In: Bewley, T. (Ed.), Advances in Economic Theory. Cambridge University Press, Cambridge.
Hastings, J.S., Weinstein, J.M., 2008. Information, school choice, and academic achievement: evidence from two experiments. Quarterly Journal of Economics 123, 1373–1414.

Hausman, J.A., Wise, D.A., 1979. Attrition bias in experimental and panel data: the Gary income maintenance experiment. Econometrica 47, 455–473.
Hausman, J., Wise, D. (Eds.), 1985. Social Experimentation. University of Chicago Press for National Bureau of Economic Research, Chicago, pp. 1–55.
Heckman, J.J., 1992. Randomization and social policy evaluation. In: Manski, C.F., Garfinkel, I. (Eds.), Evaluating Welfare and Training Programs. Harvard University Press, Cambridge.
Heckman, J.J., Masterov, D., 2005. The Productivity Argument for Investing in Young Children. National Bureau of Economic Research Working Paper Number 13016.
Heckman, J.J., Siegelman, P., 1993. The urban institute audit studies: their methods and findings. In: Fix, M., Struyk, R. (Eds.), Clear and Convincing Evidence: Measurement of Discrimination in America. The Urban Institute Press, Washington, DC.
Heckman, J.J., Smith, J.A., 1995. Assessing the case for social experiments. Journal of Economic Perspectives 9, 85–110.
Heckman, J.J., Moon, S.H., Pinto, R., Savelyev, P.A., Yavitz, A., 2010. The rate of return to the high/scope Perry preschool program. Journal of Public Economics 94, 114–128.
Heckman, J.J., Moon, S.H., Pinto, R., Savelyev, P., Yavitz, A., 2010. Analyzing social experiments as implemented: a reexamination of the evidence from the high scope Perry preschool program. Quantitaive Economics 1 (1) (forthcoming).
Hitsch, G., Hortacsu, A., Ariely, D., 2010. Matching and sorting in online dating. American Economic Review 100, 130–163.
Hoddinott, J., Haddad, L., 1995. Does female income share influence household expenditures? Evidence from Côte d'Ivoire. Oxford Bulletin of Economics and Statistics 57, 77–96.
Holmstrom, B., 1982. Moral hazard in teams. Bell Journal of Economics 13, 324–340.
Holmstrom, B., Milgrom, P., 1991. Multitask principal-agent analyses: incentive contracts, asset ownership, and job design. Journal of Law, Economics and Organization 7, 24–52.
Homan, R., 1991. The Ethics of Social Research. Longman, London.
Hossain, T., List, J.A., 2009. The Behavioralist Visits the Factory: Increasing Productivity Using Simple Framing Manipulations. NBER Working Paper 15623.
Hotz, V.J., 1992. Designing an evaluation of JTPA. In: Manski, C.F., Garfinkel, I. (Eds.), Evaluating Welfare and Training Programs. Harvard University Press, Cambridge.
Hoxby, C.M., 2000a. The effects of class size on student achievement: new evidence from natural population variation. Quarterly Journal of Economics 116, 1239–1286.
Hoxby, C.M., 2000b. Does competition among public schools benefit students and taxpayers? American Economic Review 90, 1209–1238.
Hoxby, C.M., 2000c. Peer Effects in the Classroom: Learning from Gender and Race Variation. National Bureau of Economic Research Working Paper No. 7867.
Hoxby, C.M., Muraka, S., 2009. Charter Schools in New York City: Who Enrolls and How They Affect Their Students' Achievement. National Bureau of Economic Research Working Paper No. w14852.
Hoxby, C.M., Weingarth, G., 2006. Taking Race Out of the Equation: School Reassignment and the Structure of Peer Effects. mimeo, Harvard University.
Hsieh, C.T., Urquiola, M., 2006. The effects of generalized school choice on achievement and stratification: evidence from Chile's voucher program. Journal of Public Economics 90 (8-9), 1477–1503.
Ichniowski, C., Prennushi, G., Shaw, K., 1997. The effects of human resource management practices on productivity. American Economic Review 86, 291–313.
Ichniowski, C., Shaw, K., 2008. Insider econometrics: a roadmap to estimating models of organizational performance. In: Gibbons, R., Roberts, J. (Eds.) Handbook of Organizational Economics (forthcoming).
Jacob, B.A., 2002. Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools. National Bureau of Economic Research Working Paper 8968.
Jacob, B.A., 2004. Public housing, housing vouchers, and student achievement: evidence from public housing demolitions in Chicago. American Economic Review 94, 233–258.
Jacob, B.A., 2005. Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago public schools. Journal of Public Economics 89, 761–796.
Jones, D., Kato, T., 1995. The productivity effects of employee stock–ownership plans and bonuses: evidence from Japanese panel data. American Economic Review 85, 391–414.

Jowell, R., Prescott-Clarke, P., 1970. Racial discrimination and white-collar workers in Britain. Race 11, 397–417.
Juhn, C., Murphy, K.M., Pierce, B., 1993. Wage inequality and the rise in returns to skill. Journal of Political Economy 101, 410–442.
Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decisions under risk. Econometrica 47, 313–327.
Kandel, E., Lazear, E., 1992. Peer pressure and partnerships. Journal of Political Economy 100, 801–813.
Katkar, R., Reiley, D.H., 2006. Public versus secret reserve prices in eBay auctions: results from a Pokémon field experiment. Advances in Economic Analysis and Policy 6, Article 7.
Katz, L.F., 1986. Efficiency wage theories: a partial evaluation. NBER Macroeconomics Annual 1, 235–276.
Katz, L.F., Kling, J.R., Liebman, J.B., 2001. Moving to opportunity in Boston: early results of a randomized mobility experiment. Quarterly Journal of Economics 116, 607–654.
Keane, M.P., 2010. A structural perspective on the experimentalist school. Journal of Economic Perspectives 24 (2), 47–58 (forthcoming).
Keane, M., Wolpin, K.I., 1997. The career decisions of young men. Journal of Political Economy 105, 473–522.
Kluger, A.N., Denisi, A.S., 1996. Effects of feedback intervention on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychological Bulletin 199, 254–284.
Knoeber, C., Thurman, W., 1994. Testing the theory of tournaments: an empirical analysis of broiler production. Journal of Labor Economics 12, 155–179.
Kramer, M.S., Shapiro, S.H., 1984. Scientific challenges in the application of randomised trials. Journal of the American Medical Association 252, 2739–2745.
Kremer, M., 2003. Randomized evaluations of educational programs in developing countries: some lessons. American Economic Review 90, 102–106.
Kremer, M., 2006. Expanding educational opportunity on a budget: lessons from randomized evaluations. In: Improving Education Through Assessment, Innovation, and Evaluation, American Academy of Arts and Sciences: Project on Universal Basic and Secondary Education. MIT Press, Cambridge.
Kremer, M., Moulin, S., Namunyu, R., 2002. Unbalanced Decentralization. mimeo, Harvard University.
Kremer, M., Miguel, E., Thornton, R., 2009. Incentives to learn. Review of Economics and Statistics 91 (3), 437–456.
Kreps, D.M., 1997. The interaction between norms and economic incentives intrinsic motivation and extrinsic incentives. American Economic Review Papers and Proceedings 87, 359–364.
Krueger, A.B., 1999. Experimental estimates of education production functions. Quarterly Journal of Economics 114, 497–532.
Krueger, A.B., 2003. Economic considerations and class size. Economic Journal 113, F34–F63.
Krueger, A.B., Zhu, P., 2004. Another look at the New York City school voucher experiment. American Behavioral Scientist 47 (5), 658–698.
Kruglanski, A., 1978. Endogenous attribution and intrinsic motivation. In: Greene, D., Lepper, M.R. (Eds.), The Hidden Costs of Reward. Erlbaum Pub., Hillsdale, NJ.
Ladd, H.F., 1998. Evidence on discrimination in mortgage lending. Journal of Economic Perspectives 12, 41–62.
Lavy, V., 2002. Evaluating the effect of teachers' group performance incentives on pupil achievement. Journal of Political Economy 110, 1286–1317.
Lavy, V., 2009. Performance pay and teacher's effort, productivity, and grading ethics. American Economic Review 99, 1979–2011.
Lavy, V., Paserman, D., Schlosser, A., 2008. Inside the Black Box of Ability Peer Effect: Evidence from Variation of Low Achievers in the Classroom. National Bureau of Economic Research Working Paper No. 14415.
Lazear, E.P., 1995. Personnel Economics. MIT Press, Cambridge, Mass.
Lazear, E.P., 1989. Pay equality and industrial politics. Journal of Political Economy 87, 1261–1284.
Lazear, E.P., 2000. Performance pay and productivity. American Economic Review 90, 1346–1361.
Lazear, E.P., 2001. Educational production. Quarterly Journal of Economics 116, 777–803.
Lazear, E.P., 2005. Output-based pay: incentives or sorting? In: Polachek, S.W. (Ed.), Research in Labor Economics – Accounting for Worker Well-Being, vol. 23. pp. 1–25.

Lazear, E.P., Rosen, S., 1981. Rank order tournaments as optimum labor contracts. Journal of Political Economy 89, 841–864.
Lazear, E.P., Shaw, K.L., 2007. Personnel economics: the economist's view of human resources. Journal of Economic Perspectives 21, 91–114.
Lazear, E.P., Malmendier, U., Weberz, R.A., 2009. Sorting and Social Preferences. mimeo, Stanford University.
Ledford, G.E, Lawler, E.E., Mohrman, S.A., 1995. Reward innovations in Fortune 1000 companies. Compensation and Benefits Review 27, 76–80.
Levine, D.K., Pesendorfer, W., 2002. The Evolution of Cooperation Through Imitation. mimeo, UCLA.
Levitt, S.D., List, J.A., 2007a. What do laboratory experiments measuring social preferences reveal about the real world? Journal of Economic Perspectives 21, 153–174.
Levitt, S.D., List, J.A., 2007b. Viewpoint: on the generalizability of lab behavior to the field. Canadian Journal of Economics 40, 347–370.
Levitt, S.D., List, J.A., 2009. Field experiments in economics: the past, the present, and the future. European Economic Review 53, 1–18.
Levitt, S.D., List, J.A., 2010. Was there really a hawthorne effect at the hawthorne plant? an analysis of the original illumination experiments. American Economic Journal: Applied Economics (forthcoming).
Levitt, S.D., List, J.A., Sadoff, S., 2009. Checkmate: exploring backward induction among chess players. American Economic Review (forthcoming).
Levitt, S.D., List, J.A., Sadoff, S., 2010. The Effect of Financial Incentives on High School Achievement: Evidence from Randomized Experiments. Working Paper.
Ligon, E., 1998. Risk-sharing and information in village economies. Review of Economic Studies 65, 847–864.
List, J.A., 2001. Do explicit warnings eliminate the hypothetical bias in elicitation procedures? evidence from field auctions for sportscards. American Economic Review 91 (5), 1498–1507.
List, J.A., 2002a. Preference reversals of a different kind: the 'more is less' phenomenon. American Economic Review 92 (5), 1636–1643.
List, J.A., 2002b. Testing neoclassical competitive market theory in the field. Proceedings of the National Academy of Science 99 (24), 15827–15830.
List, J.A., 2003a. Using random *n*th price auctions to value non-market goods and services. Journal of Regulatory Economics 23, 193–205.
List, J.A., 2003b. Does market experience eliminate market anomalies? Quarterly Journal of Economics 118, 41–71.
List, J.A., 2004a. Young, selfish and male: field evidence of social preferences. Economic Journal 114, 121–149.
List, J.A., 2004b. The nature and extent of discrimination in the marketplace: evidence from the field. Quarterly Journal of Economics 119, 49–89.
List, J.A., 2004c. Neoclassical theory versus prospect theory: evidence from the marketplace. Econometrica 72, 615–625.
List, J.A., 2006. Field experiments: a bridge between lab and naturally occurring data. Advances in Economic Analysis and Policy 6, Article 8.
List, J.A., Livingstone, J., 2010. Exploring the Link between Market Power and the Nature and Magnitude of Discrimination. Working Paper, University of Chicago.
List, J.A., Price, M.K., 2005. Conspiracies and secret price discounts in the marketplace: evidence from field experiments. Rand Journal of Economics 36, 700–717.
List, J.A., Reiley, D.R., 2008. Field experiments. In: Durlauf, Steven N., Blume, Lawrence E. (Eds.), The New Palgrave Dictionary of Economics, second ed., Palgrave Macmillan.
List, J.A., Sadoff, S., Wagner, M., 2010. So You Want to Run an Experiment, Now What? Some Simple Rules of Thumb for Optimal Experimental Design. NBER Working Paper No. 15701.
List, J.A., Millimet, D.L., Fredriksson, P.G., Warren McHone, W., 2003. Effects of environmental regulations on manufacturing plant births: evidence from a propensity score matching estimator. Review of Economics and Statistics 85 (4), 944–952.
Lizzeri, A., Meyer, M., Persico, N., 2002. The Incentive Effects of Interim Performance Evaluations CARESS. Working Paper 02-09.

Loewenstein, G., 2005. Hot-cold empathy gaps and medical decision making. Health Psychology 24, 49–56.
Loewenstein, G., Schkade, D., 1999. Wouldn't it be Nice? Predicting Future Feelings. In: Kahneman, D., Diener, E., Schwartz, N. (Eds.), Well Being: The Foundations of Hedonic Psychology. Russell Sage Foundation.
Longnecker, C.O, Sims, H.P., Gioia, D.A., 1987. Behind the mask: the politics of performance appraisal. The Academy of Management Executive 1, 183–193.
Lucking-Reiley, D., 1999. Using field experiments to test equivalence between auction formats: magic on the internet. American Economic Review 89, 1063–1080.
Ludwig, J., Miller, D., 2007. Does head start improve children's life chances? Evidence from a regression discontinuity design. Quarterly Journal of Economics 122, 159–208.
Lundberg, S., Pollak, R., 2003. Efficiency in marriage. Review of Economics of the Household 1, 153–167.
Lusk, J.L., Fox, J.A., 2003. Value elicitation in laboratory and retail environments. Economics Letters 79, 27–34.
Macleod, W.B., Malcomson, J.M., 1989. Implicit contracts, incentive compatibility, and involuntary unemployment. Econometrica 57, 447–480.
Malcolmson, J., 1984. Work incentives, hierarchy, and internal labor markets. Journal of Political Economy 92, 486–507.
Malmendier, U., Tate, G., 2005. CEO overconfidence and corporate investment. Journal of Finance 60, 2661–2700.
Manser, M., 1999. Existing labor market data: current and potential research uses. In: Haltiwanger, J., Manser, M., Topel, R. (Eds.), Labor Statistics Measurement Issues. The University of Chicago Press, Chicago.
Manser, M., Brown, M., 1980. Marriage and household decision-making: a bargaining analysis. International Economic Review 21, 31–44.
Manski, C.F., 1995. Learning About Social Programs from Experiments with Random Assignment of Treatments. University of Wisconsin-Madison: Institute for Research on Poverty, Discussion Paper 1061-95.
Manski, C.F., Garfinkel, I., 1992. Introduction. In: Manski, C.F., Garfinkel, I. (Eds.), Evaluating Welfare and Training Programs. Harvard University Press, Cambridge.
Mas, A., 2006. Pay, reference points, and police performance. Quarterly Journal of Economics 121, 783–821.
Mas, A., Moretti, E., 2009. Peers at work. American Economic Review 99, 112–145.
Maxfield, M., Schirm, S., Rodriguez-Planas, N., 2003. The Quantum Opportunities Program Demonstration: Implementation and Short-Term Impacts. Mathematica Policy Research Report 8279–093.
Mayo, E., 1933. The Human Problems of an Industrial Civilization. Macmillan, New York.
Mazzocco, M., 2004. Saving, risk sharing, and preferences for risk. American Economic Review 94, 1169–1182.
Mazzocco, M., 2007. Household intertemporal behaviour: a collective characterization and a test of commitment. Review of Economic Studies 74, 857–895.
McElroy, M., Horney, M., 1981. Nash-bargained decisions: towards a generalization of the theory of demand. International Economic Review 22, 333–349.
Meyer, H.H., Kay, E., French, J.R., 1965. Split roles in performance appraisal. Harvard Business Review 21–29.
Meyer, B.D., 1995. Natural and quasi-natural experiments in economics. Journal of Business and Economic Statistics 13, 151–161.
Miguel, E., Kremer, M., 2004. Worms: identifying impacts on education and health in the presence of treatment externalities. Econometrica 72, 159–217.
Milgrom, P.R., 1988. Employment contracts, influence activities, and efficient organization design. Journal of Political Economy 96, 42–60.
Milgrom, P.R., Roberts, J., 1990. The efficiency of equity in organizational decision processes. American Economic Review Papers and Proceedings 80, 154–159.
Moffitt, R.A., 1981. The Negative Income Tax: Would it Discourage Work? Monthly Labor Review.
Moffitt, R.A., 1999. Econometric Methods for Labor Market Analysis. In: Handbook of Labor Economics, Elsevier (Chapter 24).
Moldovanu, B., Sela, A., Shi, X., 2007. Contests for status. Journal of Political Economy 115, 338–363.

Moretti, E., 2004. Workers' education, spillovers and productivity: evidence from plant-level production functions. American Economic Review 94, 656–690.
Muralidharan, K., Sundararaman, V., 2007. Teacher Incentives in Developing Countries: Experimental Evidence from India. mimeo, Harvard University.
Nagin, D., Rebitzer, J.B., Sanders, S., Taylor, L.J., 2002. Monitoring, motivation, and management: the determinants of opportunistic behavior in a field experiment. American Economic Review 92, 850–873.
Nalbantian, H.R., Schotter, A., 1997. Productivity under group incentives: an experimental study. American Economic Review 87, 314–341.
Nalebuff, B.J., Stiglitz, J.E., 1983. Prizes and incentives: toward a general theory of compensation and competition. Bell Journal of Economics 14, 21–43.
Nechyba, T., 2000. Mobility, targeting and private school vouchers. American Economic Review 90 (1), 130–146.
Niederle, M., Vesterlund, L., 2007. Do women shy away from competition? Do men compete too much? Quarterly Journal of Economics 122, 1067–1101.
Olken, B.A., 2007. Monitoring corruption: evidence from a field experiment in Indonesia. Journal of Political Economy 115, 200–249.
Orne, M.T., 1962. On the social psychological experiment: with particular reference to demand characteristics and their implications. American Psychologist 17, 776–783.
Orcutt, G.H., Orcutt, A.G., 1968. Incentive and disincentive experimentation for income maintenance policy purposes. American Economic Review 58, 754–773.
Oyer, P., Schaefer, S., 2004. Why do some firms give stock options to all employees? An empirical examination of alternative theories. Journal of Financial Economics 76, 99–133.
Oyer, P., Schaefer, S., 2011. Personnel economics: hiring and incentives. In: New Developments and Research on Labor Markets, first ed., In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 4B. Elsevier, pp. 1769–1823 (Chapter 20).
Paarsch, H.J., Shearer, B., 1999. The response of worker effort to piece rates: evidence from the British Columbia tree-planting industry. Journal of Human Resources 643–667.
Paarsch, H.J., Shearer, B., 2000. Piece rates, fixed wages, and incentive effects: statistical evidence from payroll records. International Economic Review 41, 59–92.
Pahl, J.M., 1983. The allocation of money and the structuring of inequality within marriage. The Sociological Review 31, 237–262.
Peterson, P., Howell, W., Wolf, P., Campbell, D., 2003. School vouchers: results from randomized experiments. In: The Economics of School Choice. University of Chicago Press, pp. 107–144.
Pfeffer, J., 1996. Competitive Advantage Through People: Unleashing The Power of The Work Force. Harvard Business Press, Cambridge, Mass.
Phelps, E., 1972. The statistical theory of racism and sexism. American Economic Review LXII, 659–661.
Pigou, A.C., 1920. The Economics of Welfare.
Plug, E., Vijverberg, W., 2003. Schooling, family background, and adoption: is it nature or is it nurture. Journal of Political Economy 111, 611–641.
Podolny, J.M., Baron, J.N., 1997. Resources and relationships: social networks and mobility in the workplace. American Sociological Review 62, 673–693.
Prendergast, C., 1999. The provision of incentives in firms. Journal of Economic Literature 37, 7–63.
Prendergast, C., 2001. Selection and Oversight in the Public Sector with the Los Angeles Police Department as an Example. National Bureau of Economic Research Working Paper No. 8664.
Prendergast, C., Topel, R.H., 1996. Favoritism in organizations. Journal of Political Economy 104, 958–978.
Puma, M., Burstein, N., Merrell, K., Silverstein, G., 1990. Evaluation of the Food Stamp Employment and Training Program: Final Report, Bethesda, Md. Abt Associates, Bethesda, MD.
Punch, M., 1985. The Politics and Ethics of Fieldwork. Sage, London.
Raaum, O., Torp, H., 1993. AMO-kurs: Hvem søker, hvem får plass - og hvem får jobbetterpå? Søkelys på arbeidsmarkedet.
Rabin, M., Schrag, J.L., 1999. First impressions matter: a model of confirmatory bias. Quarterly Journal of Economics 114, 37–82.
Rangel, M.A., 2006. Alimony rights and intrahousehold allocation of resources: evidence from Brazil. Economic Journal 116, 627–658.

Rasul, I., 2008. Household bargaining over fertility: theory and evidence from Malaysia. Journal of Development Economics 86, 215–241.
Reback, R., 2005. School Accountability and the Distribution of Student Achievement. mimeo, Columbia University.
Rege, M., Telle, K., 2004. The impact of social approval and framing on cooperation in public good situations. Journal of Public Economics 88, 1625–1644.
Riach, P.A., Rich, J., 2002. Field experiments of discrimination in the market place. Economic Journal 112, F480–F518.
Riach, P.A., Rich, J., 2006. An experimental investigation of sexual discrimination in hiring in the English labor market. B. E. Journal of Economic Analysis & Policy 6, Advances Article 1.
Rice, B., 1982. The Hawthorne effect: persistence of a flawed theory. Psychology Today 16, 71–74.
Rivkin, S.G., Hanushek, E.A., Kain, J.F., 2005. Teachers, schools, and academic achievement. Econometrica 73, 417–459.
Robinson, J., 2008. Limited Insurance Within the Household: Evidence from a Field Experiment in Western Kenya. mimeom, UC Santa Cruz.
Rockoff, J., 2009. Field experiments in class size from the early twentieth century. Journal of Economic Perspectives 23, 211–230.
Roethlisberger, F.J., Dickson, W., 1939. Management and the Worker. Harvard University Press, Cambridge.
Rosen, S., 1982. Authority, control, and the distribution of earnings. Bell Journal of Economics 13, 311–323.
Rosen, S., 1986. The theory of equalizing differences. In: Ashenfelter, Orley, Layard, Richard (Eds.), Handbook of Labor Economics, vol. 1. North-Holland, Amsterdam.
Rosenbaum, P., Rubin, D., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.
Rosenzweig, M.R, Wolpin, K.I., 2000. Natural 'natural experiments' in economics. Journal of Economic Literature 38, 827–874.
Ross, H.L., 1970. An Experimental Study of the Negative Income Tax, Child Welfare, December.
Rotemberg, J.J., 1994. Human relations in the workplace. Journal of Political Economy 102, 684–717.
Rouse, C., 1998. Private school vouchers and student achievement: an evaluation of the Milwaukee parental choice program. Quarterly Journal of Economics 133 (2), 553–602.
Roy, D., 1952. Quota Restriction and Goldbricking in a Machine Shop. American Journal of Sociology 57, 427–442.
Rozan, A., Strenger, A., Willinger, M., 2004. Willingness-to-pay for food safety: an experimental investigation of quality certification on bidding behavior. European Review of Agricultural Economics 31, 409–425.
Rubin, D.B., 1990. Comment: Neyman (1923) and causal inference in experiments and observational studies. Statistical Science 5, 472–480.
Samuelson, P.A., 1956. Social indifference curves. Quarterly Journal of Economics 70, 1–22.
Samuelson, W., Zeckhauser, R., 1988. Status quo bias in decision making. Journal of Risk and Uncertainty 1, 7–59.
Sausgruber, R., 2009. A note on peer effects between teams. Experimental Economics 12, 193–201.
Schultz, T.P., 2004. School subsidies for the poor: evaluating the Mexican *Progresa* poverty program. Journal of Development Economics 74, 199–250.
Seabright, P., 2002. Blood, Bribes, and the Crowding-out of Altruism by Financial Incentives. mimeo, Toulouse University.
Sen, A., 1999. Development as Freedom. Knopf, New York.
Sethi, R., Somanathan, E., 1999. Preference Evolution and Reciprocity. mimeo, University of Michigan.
Shapiro, C., Stiglitz, J.E., 1984. Equilibrium unemployment as a worker discipline device. American Economic Review 74, 433–444.
Shearer, B.S., 2004. Piece rates, fixed wages and incentives: evidence from a field experiment. Review of Economic Studies 71, 513–534.
Slovic, P., Lichtenstein, S., 1971. Comparison of Bayesian and regression approaches to the study of information processing in judgment. Organizational Behavior and Human Performance 6, 649–744.
Snow, C.E., 1927. Research on Industrial Illumination. The Tech Engineering News November, 257–282.

Sobel, J., Takahashi, I., 1983. A multi-stage model of bargaining. Review of Economic Studies 50, 411–426.
Soetevent, A.R., 2005. Anonymity in giving in a natural context – a field experiment in 30 churches. Journal of Public Economics 89, 2301–2323.
Splawa-Neyman, J., 1923a. On the application of probability theory to agricultural experiments. Essay on principles, Section 9. Statistical Science 5, 465–472; Translated and edited bv D. M. Dabrowska and T. P. Speed from the Polish original, which appeared in Roczniki Nauk Rolniczyc, Tom X (1923): 1-51 (Annals of Agricultural Sciences).
Splawa-Neyman, J., 1923b. Contributions of the theory of small samples drawn from a finite population. Biometrika 17, 472–479; The note on this republication reads, These results with others were originally published in La Revue Mensuelle de Statistique, Publ. Parl'office Central de Statistique de la Republique Polonaise, Tom. vi. pp. 1–29, 1923.
Stanley, M., 2003. College education and the mid-century GI bills. Quarterly Journal of Economics 118, 671–708.
Stafford, P., 1986. Forestalling the demise of empirical economics: the role of microdata in labor economics research. In: Ashenfelter, O., Layard, R. (Eds.), Handbook of Labor Economics, Elsevier (Chapter 7).
Stevenson, B., Wolfers, J., 2007. Marriage and divorce: changes and their driving forces. Journal of Economic Perspectives 21, 27–52.
Stiglitz, J.E., 1975. Incentives, risk, and information: notes towards a theory of hierarchy. Bell Journal of Economics 6, 552–579.
Street, D., 1990. Fisher's contributions to agricultural statistics. Biometrics 46, 937–945.
Taber, C., Weinberg, B.A., 2008. Labour economics (new perspectives). In: Durlauf, Steven N., Blume, Lawrence E. (Eds.), The New Palgrave Dictionary of Economics, second ed., Palgrave Macmillan.
Thaler, R., 1980. Toward a positive theory of consumer choice. Journal of Economic Behavior and Organization 1, 39–60.
Thomas, D.A., 1990. The impact of race on managers' experiences of developmental relationships. Journal of Organizational Behavior 11, 479–492.
Thomas, D.A., 1994. Like father, like son or like mother, like daughter: parental education and child health. Journal of Human Resources 29, 950–989.
Thorndike, E.L., 1913. Educational Psychology. Oxford: Columbia University.
Todd, P., Wolpin, K.I., 2006. Assessing the impact of a school subsidy program in Mexico: using a social experiment to validate a dynamic behavioral model of child schooling and fertility. American Economic Review 96, 1384–1417.
Tsui, A.S., O'Reilly, C.A., 1989. Beyond simple demographic effects: the importance of relational demography in superior-subordinate dyads. Academy of Management Journal 32, 402–423.
Udry, C., 1994. Risk and insurance in a rural credit market: an empirical investigation in northern Nigeria. Review of Economic Studies 61, 495–526.
Ulph, D.T., 1988. A General Non-cooperative Nash Model of Household Consumption Behaviour. University of Bristol Working Paper 88/205.
Vandegrift, D., Yavas, A., Brown, P.M., 2007. Incentive effects and overcrowding in tournaments: an experimental analysis. Experimental Economics 10, 345–368.
Van Den Steen, E., 2004. Rational overoptimism (and other biases). American Economic Review 94, 1141–1151.
Veblen, T., 1934. The Theory of the Leisure Class: An Economic Study of Institutions. Modern Library, New York.
Vollmann, J., Winau, R., 1996. Informed consent in human experimentation before the Nuremberg code. British Medical Journal 313, 1445–1449.
Vollmeyer, R., Rheinberg, F., 2005. A surprising effect of feedback on learning. Learning and Instruction 15, 589–602.
Weber, R.A., 2006. Managing growth to achieve efficient coordination in large groups. American Economic Review 96, 114–126.
Weichselbaumer, D., 2003. Sexual orientation discrimination in hiring. Labour Economics 10, 629–642.
Wesolowski, M.A., Mossholder, K.W., 1997. Relational demography in supervisor-subordinate dyads: impact on subordinate job satisfaction, burnout, and perceived procedural justice. Journal of Organizational Behavior 18, 351–362.

Williams, K.T., O'Reilly, C.A., 1998. Demography and diversity in organizations: a review of 40 years of research. Research in Organizational Behavior 20, 77–140.
White, M., Lakey, J., 1992. Restart effect: does active labour market policy reduce unemployment? Policy Studies Institute.
Wuchty, S., Jones, B.E., Uzzi, B., 2007. The increasing dominance of teams in production of knowledge. Science 316, 1036–1039.
Yinger, J., 1998. Evidence on discrimination in consumer markets. Journal of Economic Perspectives XII, 23–40.
Zimmerman, D.J., 2003. Peer effects in academic outcomes: evidence from a natural experiment. Review of Economics and Statistics 85, 9–23.

CHAPTER 3

Lab Labor: What Can Labor Economists Learn from the Lab?☆

Gary Charness, Peter Kuhn
University of California, Santa Barbara, United States

Contents

☆ The authors thank Michael Kuhn and Loni Spilberg for excellent research assistance. While have done our utmost to identify as many lab experimental papers in labor economics as possible, omissions are inevitable given the volume of excellent recent research in this area. The authors regret all such omissions but nevertheless hope that our review provides a helpful guide to the "lab labor" literature to date.

ISSN 0169-7218, DOI 10.1016/S0169-7218(11)00409-6

Abstract

This chapter surveys the contributions of laboratory experiments to labor economics. We begin with a discussion of methodological issues: when (and why) is a lab experiment the best approach; how do laboratory experiments compare to field experiments; and what are the main design issues? We then summarize the substantive contributions of laboratory experiments to our understanding of principal-agent interactions, social preferences, union-firm bargaining, arbitration, gender differentials, discrimination, job search, and labor markets more generally.

JEL classification: C9; J0

Keywords: Laboratory experiment; Social preferences; Principal-agent models; Personnel economics

The economics literature has witnessed an explosion of laboratory experiments in the past 20 years. Many of these experiments have focused on topics that are central to the field of labor economics, including how workers respond to various forms of compensation, and the economics of discrimination, arbitration, bargaining, and matching. In this chapter we survey the contributions of laboratory experiments to our understanding of these questions.

We begin our review with a discussion of methodological issues: First, we pose the general question of why (and more importantly when) a labor economist might want to conduct a laboratory experiment: What types of questions, if any, are laboratory experiments best suited to answer? How do laboratory experiments compare to field experiments? Next, once one has decided to conduct a laboratory experiment, how should it be designed? Here we review the main methodological decisions an experimenter typically needs to make, and the advantages and disadvantages of the various choices.

The second half of our review turns its attention to the substantive issues in labor economics that have been addressed using laboratory experiments. While these are wide-ranging, we focus our review on the set of issues that have generated probably the largest volume of experimental papers in labor economics: the effects of compensation policies on the supply of effort by workers. We do this in two parts. The first uses "traditional" principal-agent theory as a theoretical lens to derive predictions regarding the effects of incentives, and tests these predictions in the lab. Many are confirmed; at the same time a number of robust "anomalies", such as apparent gift exchange, also appear. The second part focuses specifically on the use of experiments and the development of new theoretical models of social preferences to understand these anomalies. Finally, we also provide brief guides to the laboratory literature on a number of other labor-related topics, including union-firm bargaining, arbitration, gender differentials, discrimination, and job search.

1. WHY LABORATORY EXPERIMENTS?

Why should labor economists care about laboratory experiments? After all, there are plenty of field data available for empirical tests. In addition, there have been a number of objections to lab experiments concerning issues such as a lack of realism (external validity), demand effects, and selection effects. Indeed, the laboratory is an artificial environment. On the other hand, lab experiments have some important advantages over other approaches; we begin this section with a discussion of these advantages, then move on to critiques and responses.

1.1. Advantages of laboratory experiments

Most practitioners of lab experiments would probably agree that a key advantage is the ability to control conditions more tightly than in any other context. For example, testing theory is a basic component of both the physical and social sciences, and the scientific method relies upon explicit tests of theory. While empirical data are indeed rich and abundant, they reflect a variety of environmental factors; disentangling these factors is difficult if not impossible.

Falk and Fehr (2003) provide the concrete example of testing tournament theory (e.g., Lazear and Rosen, 1981), where contestants should in equilibrium choose effort

levels to equate marginal effort with marginal gain. Since a direct empirical test of this theory must take into account "the number of workers who compete for the prize, the effort cost functions of the workers, the exact level of the prize, and the production level including the nature of the error term" (p. 400), such a test seems impossible with traditional empirical data.[1] However, all of these factors can be controlled (and systematically varied) in a laboratory experiment. In this manner, it is also feasible to study the impact of specific institutional arrangements on behavior by systematically varying them.

Of course, the same argument can be made regarding the effects of workers' outside options, minimum wages, sick pay, discrimination, etc. As another example, the gift-exchange game (an experimental model of Akerlof (1982)) tests for a positive relationship between wages and effort; as this is a critical assumption of efficiency-wage theories (as discussed for example in Akerlof (1984) and Akerlof and Yellen (1986)), it is useful to test it; however, it is quite difficult to do so with standard field data. In general, the laboratory offers superior control, which make it possible to identify causal relationships. In field data, variables are often determined endogenously and usually it is only possible to identify correlation. Finally, if one has doubts concerning data reported in a laboratory experiment, one can readily replicate the experiment (particularly when a standard subject pool is used).[2]

One of the greatest strengths of lab experiments is the ability to take a specific theoretical model (say of behavior under a specific group incentive scheme, with no communication between players and one-shot interaction), where theory says exactly what, say, the Perfect Bayesian equilibrium should be, and have real agents play exactly that game with real monetary consequences. One can then compare the predictions of the model to what happens; if the theory is rejected it is then relatively easy to test competing explanations (e.g. inequity aversion, "reciprocity", loss aversion, framing) for the rejection. One way to think about a lab experiment is as a first link in a longer chain running from theory to actual interactions in real firms. A distinct role of lab experiments applies to situations where 'standard' game theory doesn't give us crisp predictions, as in games with multiple equilibria. Important examples include production in teams with complementarities among the agents' efforts (Brandts and Cooper, 2007), where the multiple equilibria are due to coordination problems, and repeated principal-agent interactions (e.g. Brown et al., 2004) where the folk theorem can generate an infinity of equilibria. In these cases, experiments provide information on how people behave in situations where existing theory provides little or no guide to what should happen.

The theory of mechanism design often provides suggestions for a number of mechanisms that are predicted to yield socially-efficient outcomes, or sometimes for a number of alternative mechanisms that are all predicted to be equivalent (for example,

[1] Furthermore, Falk and Fehr (2003) point out that other factors such as communication, peer pressure, and whether the interaction is one-shot or repeated may well affect behavior and are unlikely to be known to a researcher using field data.

[2] See Charness (2010) for more discussion regarding replicability.

it is well known that a properly-designed tournament should be able to exactly mimic an efficient individual piece rate). Lab experiments however can be used to show that mechanisms/institutions that work in theory don't always do so in practice, *even under conditions designed to be ideal* for the institution. Also, institutions that are predicted to yield identical results may not do so. In fact, the use of lab experiments to 'pre-test' proposed allocation mechanisms before implementing them in the real world already has an established history (see Plott (1987) for some examples). The same is true of lab experiments in at least two labor contexts: the design of matching mechanisms in professional labor markets (e.g. McKinney et al., 2005) and the design of arbitration mechanisms for public-sector union bargaining (e.g. Deck et al., 2007).

A sometimes-overlooked advantage of lab experiments is their low cost, especially compared to field experiments and survey data collection: competing explanations can often be tested or distinguished quickly and inexpensively with a modest number of sessions. In this sense, dropping laboratory experiments from our toolkit would be a little like dropping animal studies from cancer research: while results from animal studies do not always apply to humans, the ability to test many hypotheses cheaply under carefully controlled conditions provides an indispensable tool for the development of models that work in the real world.

Another comparative advantage of lab experiments is in the study of phenomena that are hard to measure in the field because they are illegal or face disapproval, such as acts of sabotage, discrimination, and spite. It is also relatively easy to measure agents' *beliefs* in the lab, using monetary incentives. This is important in view of the role played by beliefs in many game-theoretic models. Belief data from experiments has been central in the development of new behavioral theory such as guilt aversion (see Charness and Dufwenberg (2006) and Battigalli and Dufwenberg (2007, 2009)).

Finally, lab experiments offer unique opportunities to researchers who are interested in the *form of strategies* used by agents in solving dynamic problems (or playing dynamic games). An illustrative example here is the classic problem of search from a fixed wage distribution, which is often used by labor economists to model individual workers' unemployment spells. Theory has strong predictions here—that the optimal strategy has a reservation wage property—but it is difficult to test this prediction from field data (whether experimental or not) because strategies need to be inferred from choice histories. In contrast, the lab makes it easier to elicit subjects' strategies more directly in a number of ways, including asking subjects to describe their strategy (Hey, 1982), observing subjects' use of information boards (Sonnemans, 1998), and forcing subjects to play the game using the strategy method (Sonnemans, 1998; Brown et al., forthcoming). This approach has identified some interesting deviations between actual and predicted strategies (for example, subjects seem to condition their acceptance behavior on factors like their total earnings to date, which is not optimal) that researchers are now attempting to understand using a variety of behavioral approaches.

1.2. Objections to laboratory experiments and some responses

The most common objection to the data obtained through laboratory experiments is that they have nothing to do with the field environment (no external validity). In principle, this is a serious objection, of course. It comprises a number of facets, such as the fact the participants are usually undergraduate students, who typically have little experience in labor markets (particularly as firms), the issue that the stakes are low, and the fact that the "labor task" is often simply the choice of how much money to assign to another party. One might also be concerned that participants are affected by the mere fact that they are being observed (Hawthorne effects).

There are at least two main responses to these objections. First, as pointed out by Falk and Heckman (2009), "for the purpose of testing theories, [representative evidence] is not a problem because most economic models derive predictions that are independent of assumptions concerning participant pools (p. 537). Of course, it seems better to have a richer variation than is provided by undergraduate students, the most convenient source of participants for experiments conducted by academic researchers.[3] The second response involves the use of more "real-world" participants, allowing for agent self-selection *in the lab*, less artificial tasks, and higher stakes.

Laboratory experiments have been conducted on soldiers (Fehr et al., 1998), Costa Rican coffee-plantation CEOs (Fehr and List, 2004), Chinese central planners (Cooper et al., 1999), professional arbitrators (Farber and Bazerman, 1986), Ghanaian manufacturing workers (Barr and Serneels, 2009), Japanese shrimp fishermen (Carpenter and Seki, forthcoming), and employees at large French firms (Charness and Villeval, 2009), among others.[4] In fact, the performance by student participants is often fairly closely matched in such experiments. In addition, many if not most field experiments on incentive effects focus on highly specific industries or occupations, such as windshield repairers (Lazear, 2000), tree planters (Shearer, 2004), fruit pickers (Bandiera et al., 2005), and bicycle messengers (Fehr and Goette, 2007).[5] Virtually all of these seem likely to be a more highly-selected population than college students as a group, who can reasonably be considered to be representative of the college-educated labor force. Therefore, if the goal is to identify general principles that apply broadly to a large population of workers, college students might be a more attractive choice than workers in a single, narrowly-defined occupation or industry.

[3] Note that in addition to being low-cost, undergraduate participants also tend to be fairly intelligent (in fact, evidence suggests that the undergraduates who participate in experiments are more capable than the average); this is a nice combination for the experimenter.

[4] In the latter study, a mobile laboratory was taken to the workplaces of two French firms, where older (over the age of 50) and younger workers (under the age of 30) were tested in three experimental environments.

[5] A similar comment applies to "natural experiment" studies of the effects of changes in compensation policy, which have focused on stadium vendors (Oettinger, 1999), Continental Airlines employees (Knez and Simester, 2001), textile workers (Hamilton et al., 2003), steel minimill workers (Boning et al., 2007), and taxi drivers (Crawford and Meng, 2008), among others.

A second approach to the "representativeness" issue is to mimic, in the lab, the same sorts of self-selection that generate different subpopulations in the real world. Clearly, such selection can be important, for example, if altruistic workers tend to self-select into cooperative work environments (such as teams), or risk-loving (or overconfident) workers self-select into highly competitive work environments. Persons in jobs that frequently require them to make "tough" decisions such as cutting workers' pay or firing them (i.e. managers) might have highly selected social preferences indeed, so that laboratory experiments that randomly assign college students to represent "principals" may provide a particularly poor guide to the decisions of real managers. To some extent, however, laboratory experiments can allow for such self selection and even shed important light on how it works. Interesting examples of this approach in a non-labor-market context include Lazear et al. (2006), and DellaVigna et al. (2009), who allow experimental subjects to self-select out of a situation where they are "expected" to be altruistic. In a labor market context, Eriksson et al. (2008) show that allowing risk-averse subjects to self-select out of tournaments improves tournament performance. A series of interesting experiments beginning with Gneezy et al. (2003) show that women tend to self-select out of tournaments. A key advantage of the lab in addressing these self-selection questions is the opportunity to directly measure, and control for, confounding factors such as the agent's ability at the task, her perception of her own and others' abilities, loss aversion, and risk aversion.

Concerning the objection that the labor task is abstract and artificial, there has been an increasing trend in "real-effort" experiments, in which tasks have included proofreading (Frohlich and Oppenheimer, 1992), solving puzzles (Rütstrom and Williams, 2000), mazes (e.g., Gneezy et al., 2003), anagrams (Charness and Villeval, 2009), complex optimization problems (Van Dijk et al., 2001), simple clerical tasks (Falk and Ichino, 2006; Carpenter et al., 2010) and cracking walnuts (Fahr and Irlenbusch, 2000).

Regarding the issue of small stakes, laboratory experiments have been conducted in locations where the stakes translated into more than a month's earnings (e.g., Fehr et al., 2002; Slonim and Roth, 1998), with evidence that fairness considerations still seem to apply.[6] Furthermore, it is not obvious whether stakes involving larger sums of money or the small stakes that apply to decisions people make on a daily basis are more relevant for economic purposes. It is also the case that large stakes do not necessarily lead to fewer mistakes, as in shown by Ariely et al. (2008). Finally, it is possible that participants behave differently due to scrutiny (Levitt and List, 2007). As discussed by Falk and Heckman (2009), in many laboratory experiments involving more complex decisions, this is likely to be only a minor problem.[7] And of course scrutiny can be present in the field as well, as

[6] It is also worth mentioning that Bewley (1999) finds evidence of fairness considerations in his interviews of business people, Krueger and Mas (2004) find evidence of negative reciprocity amongst Firestone workers and Mas (2008) shows that police performance suffers after unfavorable arbitration decisions.

[7] They also observe that "reanalysis of the original Hawthorne data [see Jones, 1992] shows that no Hawthorne effect was present in the Hawthorne study" (p. 537).

workers are often monitored. In any event, scrutiny can be eliminated (or systematically varied); some experimenters use double-blind techniques, where payments are placed in envelopes by monitors who have not observed the experiment, so that participants understand that the experimenter cannot know their choices.

Thus, many of the objections raised against laboratory experiments are either red herrings or can be met by taking the laboratory to the field, using "real" people (of course, students are real people as well, and they respond to the financial incentives provided), real-effort tasks, and varying the stakes.[8] So, while there are certainly issues in taking the results of laboratory experiments to the field environment, these can be ameliorated. The real value of laboratory experiments is in the enhanced opportunities for, and lower cost of carefully-controlled variation, as is required for causal knowledge rather than simple correlation. This control extends to environmental features such as institutions, payoff parameters, participant pools, the nature of the interaction among the participants (e.g., anonymity versus face-to-face; one-shot versus repeated), and even the level of scrutiny. To quote Falk and Heckman (2009): "Laboratory experiments are very powerful whenever tight control ... is essential. ... Tight control ... also allows replicability of results, which is generally more difficult with field data" (p. 537). If one wishes to perform careful tests of theory, laboratory experiments are particularly useful. Of course, none of the comments above should be taken to imply that laboratory techniques are intrinsically superior to standard empirical data or field experiments. We discuss this issue in some detail in the next subsection.

1.3. Laboratory experiments and field experiments

Is a hammer a better tool than a screwdriver (or *vice versa*)? Sometimes one needs a hammer and sometimes one needs a screwdriver. They are different tools, suited for different purposes. Claiming superiority for one tool over the other seems misplaced. This principle also applies to research methods, as each method has its own strengths and weaknesses. The idea is not a new one; Runkel and McGrath (1972) identify eight research strategies (including field studies, field experiments, and laboratory experiments), which they categorize along the two dimensions of obtrusive-unobtrusive research operations and universal-particular behavior. They state: "We cannot emphasize too strongly our belief that none of these strategies has any natural or scientific claim to greater respect from researchers than any other" (p. 89).[9]

This brings us to the current debate about the value of field experiments compared to the value of laboratory experiments. List (2008) and Levitt and List (2009) extol the value of field experiments, often mentioning the notion that these are a useful bridge between naturally-occurring environments and laboratory experiments. Levitt and List (2007)

[8] A comprehensive study by Camerer and Hogarth (1999) indicates that in many cases whether or not a laboratory participant is given financial incentives has little effect on behavior.

[9] We are indebted to Keith Murnighan for this information. For more detail, see the discussion in Murninghan (2008a).

provide a criticism of laboratory experiments, pointing out factors that are beneficial in field experiments, while pointing out a number of factors that make the interpretation of data in laboratory experiments problematic. Their main issue is the degree to which "the insights gained in the lab can be extrapolated to the world beyond" (p. 153); this is also known as external validity. They mention five factors (p. 154) that can influence behavior in the lab; in our view, the three most relevant of these involve the nature and extent of scrutiny (on which the greatest emphasis is placed in the paper), the context in which the decision is embedded, and the stakes of the game.

Falk and Heckman (2009) provide a response to these comments, and strongly emphasize the value of laboratory experiments. The thrust of their argument is that the controlled variation possible in laboratory experiments facilitates tests of theory, causal effects, and treatment effects.[10] To a certain extent, field experiments also can provide fairly good control of the environment, although rarely to the level attainable with laboratory experiments.[11] Falk and Heckman address the notion that the conditions in field experiments are more "realistic"; for example, they point out how it is unclear whether undergraduate students are less representative of the overall population than sports-card traders in their natural setting.[12]

A number of arguments regarding scrutiny are mentioned above. Indeed, there is no doubt that the mere act of scrutiny can affect behavior.[13] However, since there is also scrutiny in field environments and since the sense of scrutiny and the associated possibility of "demand effects" (where the participant acts in a manner that he or she believes reflects the experimenter's desired outcome) can be nearly eliminated,[14] this concern seems somewhat overstated. The notion that social preferences can be crowded out by large financial incentives is hardly new, as it is a feature of the Rabin (1993) model. Nevertheless, as Falk and Heckman point out, many real-life decisions involve small

[10] They state: "This control allows for the testing of precise predictions derived from game-theoretic models" (p. 636).

[11] Indeed, Levitt et al. (2009) discuss some of the limitations of field experiments, such as the difficulties with some forms of replication, and the fact that "they sometimes cannot be used to distinguish between alternative theories because the experimenter exerts less control than in the lab" (p. 1414).

[12] Holt (2007) points out that field experiments are particularly valuable when social context is critically important. However, he cautions: Although field experiments can induce a more realistic social context and environment, the cost is often a partial loss of control over incentives, over measurement of behavior, or over the ability to replicate under identical conditions" (p. 14). Regarding the independence of observations, he writes: "To the extent that social context and target demographics are important in a field experiment, each field experiment is in some sense like a data point that is specific to that combination of subjects and context unless appropriate random selection of subjects is employed" (p. 14).

[13] Charness et al. (2007) provide one example, as they find that the presence of an audience of peers with a common interest can lead to more aggressive behavior in experimental games.

[14] Echoing the discussion on Section 1.2, the standard policy in the experimental laboratory at the University of Amsterdam is to separate the payment of experimental participants from any observation of behavior, by having non-experimenters place payments in envelopes that are then passed out by the experimenter. Given that this is standard procedure, experienced participants know to expect this and are not particularly suspicious about it being utilized in any specific experiment. List et al. (2004) use another approach, in which a "randomized response" means that the experimenter cannot ascertain whether a participant made a pro-social choice.

stakes, so that it is not clear that one requires larger stakes in order to provide incentives that are meaningful enough to match the relevant field environment.

Despite the divergent views expressed in these articles and others, it is worth noting that there is indeed common ground. For example, in both Levitt and List (2007) and Falk and Heckman (2009), the authors discuss how one needs a model or theory to transport findings to new populations or environments, whether these data originate in laboratory experiments or field experiments. These articles also appear to agree that the controlled variation in the laboratory is better for careful tests of theory. Both state that there are shortcomings in both laboratory and field experiments, but that each can provide useful insights. In fact, both camps apparently agree that both forms of experimentation (as well as hybrids) can be combined to yield a better understanding of the phenomena involved. One interesting point is that two of the key players in the debate (Falk and List) have used and continue to use both laboratory and field experiments in their research.[15] In a certain sense, one wonders what the shouting is about.

Our own view is that laboratory experiments are best at testing theory and identifying treatment effects, and they can also provide useful qualitative insights. However, any assumption that the quantitative levels of behavior observed in the laboratory apply to naturally-occurring settings must be carefully considered, as the laboratory is only a model of the field environment and cannot include many details that may influence behavior.[16] Field experiments, for their part, offer promise in areas that are not readily susceptible to laboratory experimentation and generally involve a greater range of personal and demographic characteristics. Field experiments are especially valuable to the extent that they can capture more realistic behavior (particularly in settings where the participants are unaware that there is an ongoing experiment). That said, a similar level of care needs to be taken in applying quantitative estimates from some highly selected field populations (fruit pickers, bicycle messengers, tree planters, school children) to other field populations. These research methodologies are complements, not substitutes. One should use the most appropriate tool or tools for the job at hand.

2. ISSUES IN DESIGNING LABORATORY EXPERIMENTS

Suppose you have decided that a laboratory experiment is a fruitful way to address a research question. This section reviews some of the main design questions the investigator

15 Both authors of this chapter have also used both methods in their research.

16 An example from the games in Charness and Rabin (2002) may provide a useful illustration. In one treatment, a participant unilaterally made a choice between (Other, Own) payoffs of (750, 375) or (400, 400); in a second treatment, the other paired participant first faced a choice between payoffs of (550, 550) or passing the choice to the second participant, who would once again face a choice between (Other, Own) payoffs of (750, 375) or (400, 400). In the first treatment, roughly half of the population chose to sacrifice 25 units, selecting (750, 375); however, only 10% did so in the second treatment. The comparison of the two treatments qualitatively shows a strong effect, which indicates that a form of negative reciprocity is present. Nevertheless one would be naive to conclude that these quantitative levels would also be observed in the outside world.

typically needs to address. We do this in two stages: first, we consider general issues that arise in almost all laboratory experiments, not necessarily restricted to questions in labor economics. Second, we focus specifically on the design of "supply of effort" experiments, which constitute the main focus of our review of the substantive research.

2.1. General design questions

The first and most basic question is how closely to try to match the field environment. This will depend to a substantial degree on whether one is testing theory, one is trying to isolate a treatment effect, or one is trying for realism in an effort to draw conclusions about the effect of policy changes in a specific environment.[17] While one should be fairly insistent that the details of the design deal correctly with the issues involved with a test of theory, one cannot expect the experimental design to precisely match the field environment. There are typically trade-offs between parsimony and richness. A general rule is to err on the side of simplicity, but to include the central elements of the question at issue. It is fundamental that the participants understand the task at hand, and this varies inversely with the degree of complexity.

Regarding the issue of comprehension, a choice variable is the degree of examples (or even actual coaching) that will be provided. Experimental practice has to some extent evolved over the years. In earlier times, it was customary to provide neither examples nor test questions, out of concern that giving examples could introduce bias or demand effects. However, this policy runs the serious risk that participants will fail to understand some important aspects of the task. It has become nearly standard practice to at least ask participants questions about what outcome would prevail in the event of various combinations of choices; it is also customary to provide examples in the instructions. Of course there is always the possibility that some bias may be introduced due to this process. Nevertheless, one can minimize this possibility by going over every contingency; if this is not feasible due to the presence of a large number of contingencies, one can select "representative" contingencies. Whether examples are needed will depend on the complexity present in the experiment.

A closely-related issue is whether to use an abstract context in the instructions or to provide a richer context that points to the field environment in question. In some experiments (e.g., Charness and Rabin, 2002), the researchers are careful to choose completely neutral terms. The main advantage of this policy is that it may well limit bias.[18] On the other hand, many laboratory experiments explicitly label the subjects'

[17] Regarding the latter case, one example comes from industry. Hewlett–Packard was interested in the consequences of changing their minimum-advertised-price policy (see Charness and Chen, 2002). Rather than use a costly test market, laboratory experiments were conducted; it this situation, it was critical to match their retailer environment as closely as possible. As another example, if firms in the experiment are losing money, it is difficult to argue that the observed behavior will persist over time in the field.

[18] However, it is quite possible that people bring their own (uncontrolled) personal experience or habits with them to the experiment, so that a neutral context may not achieved the desired effect.

roles as "firms" or "workers" (or even more specifically as "high-ability workers", etc.) and to label the choices as "wages" or "effort", etc., even though (strictly speaking) the experiment is simply a game with no actual work performed. As with providing examples, the main advantage of a richer context is that it makes it easier for subjects to understand the game, while the primary disadvantage is that it might bring in established behavior patterns/expectations from those environments.

The question of context, framing and reference points is not innocuous. Sometimes the details of instructions given to the subjects can unwittingly cause their behavior to focus on certain outcomes. One nice example of framing comes from Liberman et al. (2004). There are two treatments, both of which feature the identical prisoner's-dilemma game. However, in one treatment, the game is labeled "The Wall St. Game", while the game is labeled "The Community Game" in the second treatment. The rate of cooperation was less than 30% in The Wall St. Game, but was over 70% in The Community Game. A second example is provided by Cooper et al. (1999), who find that providing context for Chinese central planners in a lab experiment improved their understanding of the game; however this had no effect on the students in other sessions of the experiment. More recently, Levitt et al. (2009) present evidence suggesting that even professionals (such as world-class poker players who are skilled randomizers in the field) have difficulty transferring those skills to the unfamiliar context of the laboratory.[19]

Another important design question concerns whether an experiment features multiple periods or not. In our view, the first question the investigator needs to ask here is whether the real-world situation they are interested in understanding most closely resembles (a) repeated interactions between the same decision-makers over a long and indefinite horizon, (b) repeated interactions with a clear end date, or (c) one-shot interactions. Case (a) can be mimicked in the lab by having the same subjects interact repeatedly with the last period of the experiment unknown to the subjects;[20] case (b) is straightforward to implement; case (c) can be implemented either by having a single period, or by randomly re-matching subjects between multiple periods (i.e. a "strangers" design). If the experimenter is interested in one-shot interactions, having only one period is in a sense the cleanest design, but it is also the most expensive approach to gathering experimental data. It can also be problematic if agents need some experience to actually understand the game they are playing. This leads most investigators interested in one-shot interactions (where the predictions of theory are usually the sharpest) to implement multiple periods with re-matching. Interestingly, even though

[19] However, note that Palacios-Huertas and Volij (2008) find that professional soccer players (at least in the aggregate) play a mixed strategy that corresponds remarkably closely to the equilibrium predictions in the O'Neill game, where there are four choices for each of two players, but there is no pure-strategy equilibrium.

[20] Specific approaches include rolling of a die after each period with some continuation rule or having a pre-set ending that is not divulged to the participants.

"standard" game theory predicts no repeated-game effects under these conditions,[21] behavior sometimes resembles the predictions of one-shot models more closely in the last few rounds.

"Partners" designs where agents are matched for the duration of the session are, of course, expected to yield repeated-game effects; the *predicted* effects of repeated interaction typically differ dramatically between finitely- and infinitely-repeated games (with the "folk theorem" applying to the latter case). In the latter case, experiments are less useful in testing theory than in providing some idea of what tends to happen when "standard" theory has little predictive power. A related design question is whether people are always in the same role or whether this can change from period to period. There is disagreement concerning which approach facilitates learning, but role change permits the experimenter to compare an individual's behavior across the various roles.

This leads us to the question of whether to use a "within-subjects" design or a "between-subjects" design. Labor economists' experience with field data where there is typically substantial nonrandom heterogeneity disposes them towards research designs with subject fixed effects; in the laboratory, this requires administering both the treatment(s) and the control situation to the same subjects. Experimenters are accustomed to having (both observable and unobserved) heterogeneity handled by randomization, but are highly sensitive to framing and sequencing effects. Thus one's behavior under one condition may be influenced by his exposure to other conditions (something that is usually ruled out by assumption in fixed-effects econometric models). Thus, many experimenters tend to prefer "between-subjects" designs, where each subject is exposed to one and only one treatment.

Labor economists need to be aware of this motivation for the between-subject approach. In some ways, a between-subjects design is cleaner and avoids sequencing effects (although sometimes these are a main topic of interest),[22] but it is typically less powerful and costlier to implement; on the other hand, a within-subjects approach tends to be more powerful in statistical tests, but can lead to spurious correlations. An advantage of a within-subjects design is that one can control for individual differences by letting each person serve as their own control. Some experiments combine both. One way to incorporate both approaches is (a) to vary the order of treatments in a within-subject design, then (b) use only the cross-sectional data from the first treatment as a between-subjects experiment to test the robustness of the within-subject approach.

A truly crucial issue in experimental design is the calibration of the parameters, as results can be very sensitive to parameter values and functional forms. Calibration

[21] A detail that is sometimes overlooked in designs with re-matching is that, in practice, many re-matching schemes incorporate a positive probability that an agent will be re-matched with the same person at a later point in the session. This can be important in calculating expected equilibrium behavior. Sometimes investigators choose to eliminate this possibility (thus reducing the number of rounds of data that can be collected), or to minimize it by informing subjects that they will never be matched with the same partner in the next period.

[22] One can systematically vary the sequence to test for sequencing effects.

typically involves establishing a baseline for comparisons. There is little "science" to guide one to choosing parameter values; instead, this is an art that is informed by the experimenter's intuition and experience. However, one tip is to find a calibration for which the baseline treatment's results leave room to move in either direction (a calibration that leads to a very low or very high rate in the baseline permits movement in only one direction). The researcher must also consider how to justify the choice of parameter values and functional forms.

The choice of payoff method is intimately connected to the issue of calibration. Incentives should be large enough to induce thoughtful and motivated behavior by the participants. An additional consideration is whether to pay for each period or to pay the participant for only one (or several) periods randomly-chosen at the end of the session. It is more traditional to pay for each period, but there is a definite trend towards paying for only some random subset of all periods. The latter approach avoids wealth effects (the amount already earned in a session: participants often have some form of income targets), mitigates boredom in later rounds, and avoids issues of people taking chances because they know that they have negative earnings (bankruptcy) at some point in the session and that negative earnings are uncollectible.

Finally, while it is traditional to tell a responder the choice of the paired first mover before the response, a more economical approach involves contingent payments. In this "strategy method" (Selten, 1967), the responder states an action at each and every information set. This permits the researcher to obtain an observation at every node of the game, which is particularly valuable when a node is reached rather infrequently. However, while the quantity of data is maximized, there remains the issue of the quality of the data. The strategy method is quite popular, but remains controversial. The most exhaustive study to date (Brandts and Charness, forthcoming) examines many comparisons of results with the two methods, finding that there is generally no qualitative difference; we are unaware of any experiment in which a treatment effect is found using the strategy method that vanishes when the game is played through ("direct response"). In any case, this is an arrow in the experimentalist's quiver and is something to consider.

We close this section with a list of "fatal errors" mentioned in Holt (2007, p. 14):

1. Inadequate or inappropriate incentives
2. Non-standardized instructions and procedures
3. Inappropriate context
4. Uncontrolled effects of psychological biases
5. An insufficient number of independent observations
6. Loss of control due to deception or biased terminology
7. The failure to provide a calibrated baseline treatment.
8. The change in more than one design factor at a time.

Needless to say, one should endeavor to avoid these pitfalls.

2.2. Design questions in principal-agent/effort experiments

Principal-agent experiments, discussed in detail in Sections 3 and 4, are a broad class of experiments in which a principal (who in some cases is the experimenter herself) first specifies a "contract" that describes how the agent (who moves second) will be rewarded as a function of his performance in a task. In the second stage, the agent performs the task, choosing—among other things—how much effort to expend. Principal-agent experiments are perhaps the largest class of lab experiments of interest to labor economists; this section discusses some design issues specific to these types of experiments.

A first question in these experiments is whether there will be a market for contracts. Most experiments simply start with firm-worker pairs that can realize some rents if they make an exchange and have a fixed outside option if they do not exchange. For many questions this is perfectly fine. But this leaves no room for labor markets, which can affect and be affected by the nature of principal-agent interactions. Early principal-agent experiments (e.g. Fehr et al. (FKR) 1993) incorporated an *ex ante* market for labor contracts, and showed that fairness considerations in the principal-agent interactions caused that market to fail to clear. More recently, Charness et al. (forthcoming) have shown how causation can run the other way: introducing *ex post* labor markets can eliminate the well-known ratchet effect in the repeated principal-agent problem.

If one chooses to implement a market in the laboratory, how can this be achieved? One approach involves an auction, as in FKR, who set up a two-stage game in their experiment. The first stage was a one-sided oral auction in which firms made wage proposals, but could not choose any individual worker, as every worker could accept every offer. If a worker accepted an offered wage, a binding contract ensued; people who were not paired at the end of three minutes received zero profits for this period. In the second stage, workers chose effort anonymously (only the paired firm learned the chosen effort). Other, simpler approaches to modeling agents' outside options include simply manipulating the agent's compensation if he/she chooses not to work for the principal to whom he/she has been assigned, or allowing an agent to receive simultaneous offers from more than one principal (see for example Charness et al., forthcoming).

As mentioned earlier, some labor experiments use some form of real effort, while others use a stated effort level that is simply a transfer (at some rate of exchange) from the agent to the principal. An advantage of stated effort is that we know the disutility-of-effort function and can therefore calculate exactly what the equilibrium effort levels should be, according to different theories. This approach also allows the investigator to induce, and manipulate, differences in ability/cost of effort, separately from other personal characteristics (e.g. risk aversion, competitiveness, reciprocity) that might be correlated with it in a sample of persons. Of course, the advantage of using real effort

is that the task is more in line with what most people consider labor, and so might be considered to be a better match to the field environment.[23]

Another design issue concerns who plays the role of the principal. One approach is to place all participants in the role of agents, with the agent's compensation scheme manipulated by the experimenter. If the researcher's only interest is in the response of agents to different compensation schemes, one might argue that this is the simplest and most economical design: all subjects are agents, essentially working "for" the experimenter by performing either a real task or selecting a level of "chosen effort".[24] An alternative design assigns some subjects to the role of principals, who choose compensation policies to which agent-subjects respond. Some arguments in favor of the former approach are that (a) subjects might be more disposed to treat the experimenter (as opposed to a fellow student in the lab) like a "real" employer, and (b) the behavior of college students acting as firms provides little insight into the behavior of "real employers". Also, when participants choose the pay scheme, it is *not* randomly assigned. On the other hand, the latter approach (with subjects as principals) may have advantages if one is interested in social preferences towards persons other than the experimenter, or in the behavior of principals *per se* (for example if the subjects are experienced managers).

Additional considerations in the design of principal-agent experiments include whether workers can self-select among reward schemes, such as a tournament or a piece-rate scheme (Niederle and Vesterlund, 2007); whether the experimenter induces reference points (Abeler et al., 2009); allows communication (Charness and Dufwenberg, 2006; Brandts and Cooper, 2007); allows for a monitoring/fines technology (Fehr et al., 2007); or for some coercion of agents (Falk and Kosfeld, 2006). In the case of multiple agents per principal, the experimenter needs to decide whether agents can observe each other's actions (Falk and Ichino, 2006) or wages (Charness and Kuhn, 2007), whether pay is based on relative performance, or whether subjects interact in teams. All of these, and related questions, constitute the fabric of an extensive research agenda on principal-agent interactions in the lab, which we review in detail in Sections 3 and 4 below.

2.3. Reading papers involving laboratory experiments

Labor economists may be at a loss in reading papers that report the results of laboratory experiments. The format may well be unfamiliar, the design mysterious, and the statistical methods foreign. In addition, many experimental papers seem written for experimental

[23] A related issue is the distinction between hours of work and effort. While "traditional" labor supply theory is framed as a worker's choice of hours at a fixed hourly wage, principal-agent theory and virtually all labor experiments are framed as a choice of effort. While in many ways the choices are isomorphic, there are some important distinctions; caution is required when generalizing "effort" results to choices of hours worked. See for example Dickinson (1999).

[24] Another design option is to use automated agents when there is really only one sensible response for some types of agent. In some cases this can substantially increase the amount of data that can be collected. See for example Charness et al. (forthcoming).

audiences, rather than the general population of economists. Nevertheless, there are some pointers that can be provided for labor economists interested in gleaning the substance and details of experimental papers.

Perhaps the most important factor in reading an experimental paper is to understand the experimental design. This is not always as clear as it should be in the text; often, experimental referees first read the experimental instructions. These should be consulted if there is any doubt concerning the exact procedures. It is critical for the reader to understand the flow of information; this means knowing what the participants knew and when they knew it, in terms of the stages of the experimental game or task. When the design is complex, there is also a concern that participants may not have understood the game or task involved.

Once the instructions are understood, the reader should consider how well the experimental design constitutes an appropriate test of theory or matches the "ideal" field environment of interest. While one should be fairly insistent that the details of the design deal correctly with the issues involved with a test of theory, one cannot expect the experimental design to precisely match the field environment. Still, it is important that the reader is persuaded of the relevance of the experiment to the field or to the theoretical environment. The reader should also be alert to the issue of framing, given the substantial possibility that this affects behavior. To a certain degree, framing effects may wash out when one compares across treatments, but this can be a delicate issue.

An important issue when reading an experimental paper is the presentation and analysis of the data. If one is concerned with "where the bodies are buried" some degree of caution may be sensible. Authors have been known to put the best face on the data (for example, empirical researchers may tend to report the more useful regressions), so readers should keep this in mind. For example, sometimes articles emphasize (or only report) data from a subset of the periods; at times this can be justified and at times it is convenient. Sometimes authors will pool data from treatments; this increases the number of observations and makes statistical tests more powerful, but this pooling must be justified. In general, it has happened that experimental papers (and others) have interpreted their results in a favorable light. One should consider whether these interpretations are justified and whether there are alternative interpretations.

Regarding the issue of statistical and econometric tests, since most labor experiments in the laboratory feature multiple periods and interaction amongst the participants, one must have some approach towards determining how to treat multiple observations for the same individual. Labor economists are very familiar and comfortable with panel-data techniques, but experimenters are less so. Some feel that each session can only be considered to present one independent observation. A less strict approach is to collapse each individual's choices to an average, eliminating the issue of multiple observations (but not eliminating the issue of interactions during the session). In either of these cases, it is common for experimenters to report non-parametric tests, and sometimes no regressions

are reported; labor economists may be unfamiliar with these tests. One workhorse is the Wilcoxon rank-sum test, which ranks the behavior of individual participants of individual sessions in each treatment and then compares the sums. When within-subject data are available, the binomial test is often used; here one can compare changes for each individual across tasks. If these changes go predominantly in one direction or the other, one can conclude statistically that the behavior is significant. The reader should understand how these tests (and the ones reported in the article) work.

3. TESTING "TRADITIONAL" PRINCIPAL-AGENT THEORY IN THE LAB

The question of how workers' choices of effort and work hours respond to financial incentives is among the oldest questions in labor economics. In this section we consider how these questions have been addressed in the laboratory; our treatment roughly follows the literature on principal-agent models and in personnel economics by beginning with the simplest forms of work incentives (a wage per hour worked or an individual piece rate), moving on to incentives based on relative performance (tournaments), incentives for teams, multitask settings, and multi-period principal-agent settings. Not only does "traditional" principal-agent theory serve as a useful organizing device for our discussion, many of its predictions are confirmed in the lab.[25]

That said, the experiments summarized in this section yield a number of robust results that are inconsistent with standard principal-agent models, including for example a strong apparent tendency by workers to "reciprocate" generous wage offers from firms, even when such reciprocal behavior is costly to workers. In Section 4 we focus specifically on the use of experiments and the development of new theoretical models of social preferences to understand these "anomalies", with the ultimate goal of developing a more general class of models that is more firmly grounded in empirical fact and might be dubbed behavioral principal-agent theory.

3.1. The basic principal-agent problem: One principal, one agent, one task, and one interaction

3.1.1. Animal labor supply experiments

To the best of our knowledge, the earliest economic studies of the effects of material incentives on labor supply in the laboratory were the animal experiments of the early 1980s (Battalio et al., 1981; Battalio and Kagel, 1985).[26] Much of this work is summarized in Kagel et al. (1995); see also Kagel (1987) for a general discussion of

25 For an alternative and complementary review of some of these issues, see Camerer and Weber (forthcoming).

26 Battalio et al.'s work, in turn, has roots in an extensive literature in experimental psychology on the effects of reward structures and amounts on animal behavior. See for example Ferster and Skinner (1957), Barofsky and Hurwitz (1968) and Kelsey and Allison (1976). These studies commonly find backward-bending labor supply curves as the reinforcement rate is increased, but do not relate their results to a theory of utility-maximization, or consider the effects of income-compensated changes in "wages".

the contribution of animal experiments to economics. A key objective of these studies was to test the classic, static economic model of labor supply in which an agent chooses consumption (C) and leisure (L) to maximize a quasiconcave utility function $U(C, L)$, subject to the constraint $C = wL + G$ where w is the wage rate and G is unearned income. In these experiments, hungry animals expend real effort (key pecking for pigeons, lever presses for rats) to obtain income; the experimenters then vary both parameters of the budget constraint (w and G) exogenously and study the animals' reactions.

The key prediction tested by the authors is the labor supply response to an income-compensated wage decrease. As predicted by the standard model, both pigeons and rats reduce their labor supply and consumption (Battalio et al., 1981; Battalio and Kagel, 1985). The authors also study the pure income effects of declines in nonlabor income (G): in virtually all cases these raised labor supply, indicating that leisure is a normal good. The normality of leisure means that it is common to observe backward-bending labor supply curves among animal workers (see for example Battalio et al., 1981, Table 3).[27]

Another interesting feature of the animal studies that generalizes to the plethora of human studies is the presence of large subject effects: while most subjects respond to changes in incentives in the direction predicted by simple utility maximizing models, both the level of effort at any given reward *and* its responsiveness to incentives vary widely across subjects.

3.1.2. Piece rates and effort

To the best of our knowledge, the first laboratory experiment to examine labor supply responses to wage changes among humans that is couched in economic theory appeared in an accounting journal (Swenson, 1988). Swenson's subjects supplied "real" effort (repeatedly typing "!" then "enter" on a computer keyboard—this requires two hands and does not allow for continuous cursor movement).[28] Wages per character typed were fixed, but "taxed" (this language was used in the subjects' instructions) at rates ranging from 12 to 87%. Total tax proceeds from the previous session were randomly distributed to the subjects in the following period, mimicking a balanced government budget but breaking most of the connection between current individual effort and future lump-sum income. The primary questions addressed were how labor supply and total tax revenues

[27] Other labor economics questions addressed by these researchers in the lab include the "welfare trap" (do high levels of unearned income in past treatments reduce current labor supply?), and the "cycle of poverty": does deprivation raise animal subjects' discount rates (i.e. their preference for a small immediate reward over a larger delayed one), thereby leading to more deprivation in the future? Interestingly, little evidence of either effect was found (Kagel et al., 1995).

[28] As discussed in Section 2.2, laboratory principal-agent experiments with human subjects can be divided into those where subjects are paid to perform an actual task ("real-effort") and those where effort decisions are represented by the choice of a decision number that imposes increasing marginal financial costs on the agent ("chosen effort"). Bruggen and Strobel (2007) find little difference between the two methods in a simple gift-exchange labor market game.

respond to the tax rate. Both curves were backward-bending, with tax revenues (i.e. the Laffer curve) peaking at the 73% tax rate.[29]

A decade later, economists Sillamaa (1999a,b) and Dickinson (1999) conducted similar real-effort experiments.[30] Like Swenson's, Sillamaa's experiments were motivated by questions about the impact of taxation (in one case, the impact of tax progressivity, in the other the effect of a zero top marginal rate), though in her case taxes were never mentioned in the subjects' instructions. Sillamaa found that (a) work effort responds more (positively) to real wage increases in the presence of an (equivalent) linear than a progressive income tax, and (b) introducing a zero top marginal tax rate also increased effort.

Like Sillamaa, Dickinson (1999) paid his subjects a piece rate, but in some treatments allowed his subjects to choose between two types of leisure: on- versus off-the-job. This modification is noteworthy because it provides one of the few empirical links between the types of work decisions that are usually studied in lab (and field) experiments (effort) and the traditional application of labor supply theory (to hours worked). Specifically, in the baseline ("intensity") treatments, subjects were required to stay for the entire two-hour experimental period; thus any time not working was spent in the lab. In the "combined" treatment, subjects could leave at any time during the experimental period. Consistent with theory and with previous research, subjects increased their output in the baseline treatment, substituting on-the-job leisure for effort when incentives were strengthened. In the combined treatment, many subjects responded to higher wages by working more quickly, but reducing their total work time by leaving the experiment early. This substitution of off-the-job for on-the-job leisure is offered as a possible explanation for why econometric estimates of labor supply elasticities (which use hours worked, not effort as their measure of labor supply) are often close to zero. Dickinson's analysis also points out that care must be taken in relating the results of laboratory labor supply experiments (where workers' *effort* during a fixed work period is the outcome of interest) to econometric studies of labor supply (where *hours worked* is the outcome).

Gneezy and Rustichini (2000) also studied the response of work effort to financial incentives; they conducted real-effort experiments in both the lab and the field, with similar results: the relationship between the piece rate and effort was U-shaped, with low piece rates eliciting less effort than a zero piece rate. (It may be interesting to note that this is exactly the opposite of the backward-bending labor supply curve in the one-period neoclassical model, which yields an inverted U). The authors hypothesize that small

[29] It may be worth noting that backward-bending labor supply in the context of a laboratory experiment is actually somewhat of a puzzle for humans, since, at least in the classic intertemporal labor supply model, implausibly severe liquidity constraints would be required to generate such effects for payments of the size and duration that are typical for lab experiments.

[30] In Sillamaa's experiments, workers decoded numerical codes into letters; in Dickinson's workers repeatedly typed paragraphs, with a penalty for mistakes. Sprinkle (2000) considers the effect of strengthening incentives for a considerably more complex task, spanning multiple periods and requiring belief revision and judgment calls. Here, incentives again increase effort, but only after the subjects had been exposed to considerable feedback and experience.

levels of financial compensation (explicit incentives) may "crowd out" workers' intrinsic motivation to perform these tasks.[31] While this explanation may be more relevant to their field experiment (where the workers solicited charitable contributions) than their lab experiment (which had no charitable component) the phenomenon was observed in both settings. Since earlier studies of piece rates did not, to our knowledge, implement treatments with a zero rate, Gneezy and Rustichini's results do not necessarily conflict with those findings, whether on human or animal subjects.

In a more recent real-effort experiment, Cadsby et al. (2009) show that the effect of performance incentives varies with agents' risk aversion. In their experiment, 25% of subjects actually perform worse when incentives are intensified; further the probability of such deterioration increases with risk aversion and with measures of stress. A similar result is obtained by Ariely et al. (2008), who exposed subjects in the US and India to incentives ranging from small to very large (relative to their typical levels of pay). In many cases, very high rewards had a detrimental effect on performance. Combining these results with the nonmonotonicity identified by Gneezy and Rustichini (2000) above suggests that the effect of stronger incentives on performance, predicted to be monotonic by basic labor supply theory (at least when income effects are unimportant, which is expected for laboratory experiments on humans), may in fact be highly non-monotonic.

3.1.3. Selection into piece rate compensation

Since Lazear's famous Safelite study (2000) economists have realized that a significant share of the productivity improvements associated with piece rates can take the form of voluntary self-selection of higher-productivity workers into piece rate schemes, rather than changes in the work effort of existing workers. Laboratory studies that allow for self-selection into different pay schemes abound, though many of these focus specifically on selection into tournaments and teams. These studies are discussed later in this section. A recent study that considers the self-selection that occurs when a simple piece rate is introduced is Cadsby et al. (2007). As in most studies, pay-for-performance raises productivity. Like Lazear (2000), they also find that sorting enhances this effect: more productive employees are more likely to choose pay-for-performance schemes.[32]

[31] Psychologists provided compelling evidence that monetary incentives can crowd out intrinsic motivation long before Gneezy and Rustichini's paper (see Deci et al., 1999 for a meta-analysis). See also Frey and Oberholzer-Gee (1997) for an earlier economic analysis, though not in the domain of labor economics.

[32] Dohmen and Falk (2006) examine laboratory subjects' voluntary self selection into several types of contracts (fixed pay, piece rate, tournament or revenue-sharing scheme) according to subjects' risk attitudes, overconfidence, social preferences, gender and personality. They find that most of the extra output generated by all three variable pay schemes is due to the selection of abler workers into them. Burks et al. (2009) provide additional evidence on the importance of selection into pay-for-performance by conducting a context-rich prisoner's dilemma lab experiment on bicycle messengers who are employed in three types of firms: firms using pay for performance, those using hourly wages, or workers' cooperatives. Workers in the first type of firms were the least cooperative in prisoners' dilemma games. Using data from cities where bicycle messengers were not able to choose among these firm types, the authors argue, however, that these differences are not due to selection, but due to workers' (rapidly) adopting the norms governing cooperation in their workplace.

A somewhat different perspective on selection into pay-for-performance schemes is provided by two recent papers by Bandiera et al. (2007, 2009a) who conducted field experiments in a fruit-picking firm. When a pay-for-performance element (based on their unit's output) is added to *managers'* compensation schedules, Bandiera et al. (2007) find that managers are more likely to select able workers into the units they manage. In their 2009 paper, the same authors show that this shift towards abler workers came at the expense of workers who were socially connected to the manager. In both papers, the shift away from friends increased the work group's total output and the manager's compensation. In contrast, Belot and van de Ven (2009) find in a field experiment with children that agents who are selected because they are friends increase their subsequent performance, presumably to reciprocate the favor of being selected. In such cases, favoring one's friends can be costless, or even beneficial to the manager and the firm. To our knowledge, the effect of favoritism on selection into pay-for-performance has not yet been studied in the lab.

3.1.4. Reciprocal behavior

Evidence on the apparent presence of reciprocal behavior in workers' effort choices in economics goes back at least to the pure gift-exchange labor markets implemented by Fehr et al. (1993). Labor contracts in these settings contained no explicit incentives; despite this, workers supplied costly effort, and supplied more effort the higher the (lump sum) wage the principal paid them. We treat this "pure" gift exchange literature in another section; here we provide one or two examples of how workers' apparent concerns for reciprocity in the laboratory affect the performance of standard incentive contracts, such as piece rates.[33]

An illustrative paper in this regard is Anderhub et al. (2002), who study the behavior of both principals *and* agents where the contract specifies the agent's pay as a linear function of his/her output. (All the papers considered thus far study agents' reactions to reward schedules set by the experimenter.) Because there is no uncertainty, the efficient linear contract has a piece rate of 100%. Further, because principals make take-it-or-leave-it contract offers to agents in this experiment, the intercept term of the equilibrium linear contract is predicted to extract all of the agent's surplus if social preferences are absent. As one might expect, principals and agents behave relatively efficiently with respect to the slope of the contract (principals choose a 100% piece rate 30% of the time and a positive piece rate 98% of the time; agents optimized against this, choosing conditionally rational effort levels 87% of the time). Social preferences, however, clearly affected both the principals' choice of the intercept and agents' responses to it: agents rejected offers that split the surplus too unevenly, and principals made few such offers. There was also

[33] Our section on the role of reciprocity and social preferences in principal-agent interactions also considers the closely-related role of communication, especially its ability to foster trust, cooperation and guilt. See for example Charness and Dufwenberg (2006, forthcoming).

some tendency for generous offers to lead to higher effort levels, though as already noted the vast majority of effort decisions were egoistically rational given the piece rate.

While effects such as those reported above are both dramatic and common in lab experiments, we note that more recently, Gneezy and List (2006) have argued that positive reciprocity effects detected in lab experiments can wear off very quickly in the field; Kube et al. (2006a,b) in turn generate longer-term effects of reciprocity in the field, especially for negative reciprocity. We discuss these questions further in the section on reciprocity and social preferences. Finally, we note that, in addition to modifying the nature of principal-agent interactions, social preferences may also explain why some principal-agent relationships exist in the first place. For example, Hamman et al. (2008) report on an experiment in which principals can hire agents to behave selfishly on the principal's behalf. Delegation of decisions that would otherwise make the principal act directly in a selfish manner appears to yield more lucrative outcomes for principals.[34]

3.1.5. Reference points

An emerging issue in the study of one-on-one principal-agent relationships is the effect of reference points on effort provision. Part of the inspiration for this is a lively debate in the non-experimental literature on the presence of reference points in labor supply decisions by agents (in particular, taxi drivers and bicycle messengers) who can vary their hours and effort on a daily basis (see for example Camerer et al., 1997, Farber, 2005, Fehr and Goette, 2007, Farber, 2008, Crawford and Meng, 2008). One advantage of addressing this issue in the lab is that some possible reference points—for example, expected earnings in a round or session—can not only be observed, but manipulated by the experimenter. This is the approach taken by Abeler et al. (2009).

In Abeler et al.'s experiment, subjects are paid a piece rate to perform a tedious task. At the end of the period, with 50% probability they are paid their accumulated piece rate; otherwise they receive a fixed payment that is known in advance. Subjects decide how much to work before they know whether they will receive the fixed payment or their accumulated piece-rate earnings. Abeler et al. find significant bunching of piece-rate earnings at the level of the fixed payment. Further, this spike in the earnings distribution moves when the fixed payment is changed. Neither of these is consistent with the "standard" effort-leisure choice model (unless one were to introduce fairly unusual forms of non-separability between income and leisure). Instead, the authors argue that their results are consistent with Koszegi and Rabin's (2006) model of reference-dependent preferences, where the reference point is the subject's expected earnings for the experimental session (which is manipulated by the experimenter). The authors take considerable care to ensure that subjects' choices of "target" earnings are not driven

[34] See also Bartling and Fischbacher (2008) and Coffman (2009).

purely by the salience of those particular numbers in the instructions and experimental environment.[35]

3.1.6. Motivational 'crowding out'

Another question addressed in the experimental literature on worker-firm interactions is the effect of certain "coercive" features of contracts, such as minimum effort requirements or employee monitoring, on agents' effort levels (and more broadly, on contract efficiency). In this regard, Frey (1993) proposed that, especially in environments where the principal and agent know one another personally, the principal's decision to monitor the agent may be interpreted as a signal of distrust, and may reduce effort despite the obvious direct "disciplining" effects of monitoring. Falk and Kosfeld (2006) test a closely related idea in the lab using a very simple game where agents choose effort (which costs them less than it benefits the principal), and principals' *only* decision is whether to impose a minimum effort level on the agent. (This is essentially a gift-exchange game without an initial "gift"—agents' endowments are positive and principals' are zero.) If principals' decisions to impose a minimum effort level had no effects on agents' behavior, the truncated distributions of agents' effort (above the imposed minimum) should be the same whether the minimum is imposed or not. This is not the case: Falk and Kosfeld find "hidden costs of control" in the sense that the majority of agents reduce effort when firms attempt to "control" their actions (though the effort levels of a smaller number of "opportunistic" agents were mechanically increased by the effort minimum). In most treatments, these net reductions in effort were so substantial that principals who "controlled" earned lower payoffs then those who did not. In a follow-up survey, the authors asked agents the free-form question "What do you feel if [the principal] forces you to transfer at least [x] points?" The most common response was "distrust", especially among agents who reacted negatively to control. The authors' results suggest that, at least in a laboratory environment, rigid attempts to control agents' behavior can "backfire"; the authors also provide some support for the external validity of their results by administering a survey eliciting students' self-reported "work motivation" in a variety of hypothetical work situations involving different degrees of employer control or trust.

In a clever variation on Falk and Kosfeld's design, Schnedler and Vadovic (2007) show that control by principals does not elicit negative reactions from agents when the principal's control is legitimized in two alternative ways. In one of these, the principal must set a common control policy that applies not only to the agent, but also to a computerized "automaton" agent who supplies minimum effort whenever this is allowed. Perhaps not surprisingly, agents "understand" the principal's decision in this case and do not reduce their effort when controls are imposed. In the other, the principal is given a small endowment (in contrast to zero in Falk-Kosfeld), and agents are allowed to

35 Another possible reference point in a bilateral relationship is the terms of an *ex-ante* competitively negotiated contract between the parties. Experimental evidence of these effects is provided in Fehr et al. (2008), though not in a labor or principal-agent context.

take from this endowment by choosing a very low effort level. Here as well, experimental subjects treat control decisions that simply protect the principal's endowment from agent "pilfering" as legitimate.

Irlenbusch and Sliwka (2005b) suggest an intriguing explanation of the negative incentive effects of paying for performance in a simple experiment where principals and agents first play a pure gift-exchange game, followed by a game in which principals had the option of offering a piece rate in addition to the fixed payment (gift). Consistent with Gneezy and Rustichini (2000), agents' effort actually fell after a low piece rate was introduced. This is particularly interesting since the task performed by the agents was chosen to yield little or no intrinsic reward. More importantly, effort fell even further when, in a third treatment, piece rates were once again disallowed. A possible explanation is that agents' perception of the implicit contract offered by principals is changed by the introduction of piece rates: the presence of piece rates signals that agents are expected to behave egoistically; while the offer of a fixed wage signals that, as in many real-world employment relationships, a reasonable amount of effort is simply expected in return for a wage. Certainly, Irlenbusch and Sliwka's results suggest that *studies of what appears to be intrinsic motivation should pay close attention to subjects' interpretation of the implicit contractual understandings that may be signaled by different pay schemes*.

Fehr et al. (2007b) find a similar pattern when they compare the performance of three types of contracts in a simple laboratory experiment. In an "incentive" contract, the principal stipulates a wage w, a required effort level e^*, and a fine f. If the agent accepts the contract, he is 'audited' with exogenous probability p, and is forced to pay the fine if the effort he has chosen falls short of e^*. In a "bonus" contract, the principal announces a wage, a *desired* effort e^* and her (unenforceable) intention to pay a bonus b if $e \geq e^*$. Finally, a "trust" contract is pure gift exchange in which the principal offers a wage and simply requests effort in return.[36] For their parameterization, FKS find that, when principals must choose between trust and incentive contracts, incentive contracts performed better: they yielded higher effort levels, higher payoffs for both principals and agents, and were increasingly selected by principals over the course of the experiment. These results are consistent with findings of Lazear (2000) and others that incentives increase effort. In contrast, however, when principals must choose between incentive and bonus contracts,[37] bonus contracts dominate incentive contracts: they constitute the overwhelming majority of contracts offered, and yield higher levels of effort and payoffs to principals; this result contradicts the predictions of contract theory with egoistic agents. The authors explain these contrasting results by parameterizing Fehr and Schmidt's (1999) model of inequity aversion. Essentially, if the 'fair-minded' share of the population is neither too high nor too low, there are too few fair-minded persons to

[36] Note that the agent's actual effort, e, is observed by the principal *ex post* in all contracts, whether the agent is "audited" or not; auditing simply means that a punishment can be enforced if effort is short of the requirement.

[37] Note that the class of bonus contracts includes the "trust" contract as a special case, with bonus of zero.

make trust contracts perform best, and two many fair-minded persons to make incentive contracts work best. That said, the authors recognize that inequity-aversion is not the only possible explanation for their results; indeed, the likelihood that incentive contracts signal distrust (plus the fact that the authors constrain the enforcement technology in incentive contracts to make a first-best allocation infeasible) may also help explain this pattern of outcomes.

Dickinson and Villeval (2008) also consider the effect of monitoring (the key element of FKR's "incentive contract") on work effort; their setting is a real effort laboratory experiment, where the task was designed to contain an element of intrinsic motivation.[38] Principals choose monitoring intensity, which raises the probability the agent is audited (and penalized via a "fine" paid to the principal if his output did not exceed the target). Dickinson and Villeval vary two main aspects of the environment: in the "variable" treatment, the principal's profit, as usual, depends directly on the effort chosen by the agent. In the "fixed" treatment it does not. The other aspect that is varied is the degree of anonymity; interaction is either anonymous or preceded by five minutes of face-to-face interaction. Dickinson and Villeval find that monitoring raises agents' effort in the anonymous setting, as predicted in the standard agency model. Motivational crowding-out is observed *only* when interactions are not anonymous *and* when the principal's payoff depends directly on the agent's effort (their "variable" treatment). This suggests that the *motivational "crowding out" by monitoring is not driven primarily by a reduction in the intrinsic rewards derived from the task, but from a form of negative reciprocity (punishing the principal for a lack of trust)*.[39]

3.1.7. Nonlinearities: Targets, fines and bonuses

While most of the experimental literature on piece rates considers simple linear reward schedules, and while linear contracts are theoretically sufficient to achieve efficiency when agents are risk neutral, it is interesting to consider the effects of nonlinearities in individual piece rate contracts, since these do occur in the real world.[40] The one experimental paper we know of that focuses on this topic is Cadsby et al. (2008). The distinguishing feature of their experiment is that the task allowed agents to misrepresent their own performance (the number of words created in an anagram game). While actual output was similar under target-based pay schemes versus a continuous (linear) reward scheme, the former produced significantly more cheating. Further, cheating is more likely under a target-based scheme the closer a participant's actual production is to the

38 Agents try to attain a maximum value of a non-decreasing function by costly sequential search.

39 Ellingsen and Johannesson (2008) provide a theoretical model of crowding out in which some audiences are more worth impressing than others. In their model, the principal's choice of monetary incentives signals that he/she is not worth impressing. See Bénabou and Tirole (2003, 2006) for closely related models.

40 Of course, the voluntary bonuses studied by Fehr et al. (2007a) are nonlinear *ex post*, but here we focus on enforceable reward schedules, announced *ex ante*, that contain discrete jumps. Incidentally, Fehr et al. (2007b) adds an enforceable fine to Fehr et al. (2007a), but finds it is rarely selected by principals and has little effect on agents' effort.

target. Since the agent's rewards to cheating are also greatest in these situations, Cadsby et al.'s results are both consistent with theory and indicative of a possible drawback with sharp discontinuities in reward schedules.

3.1.8. Peer effects and wage comparisons

Although "peer effects" on effort can only exist when a firm employs multiple workers, we consider peer effects in this section on one-on-one principal-agent interactions because "pure" peer effects refer to a situation where workers work, side by side, for the same firm but do not interact in any way (except that they observe each others' work activity). For example, suppose that two workers are each paid an individual piece rate, and there are no substitutabilities or complementarities in production, but can observe each other's effort or output. Does anything change? Perhaps surprisingly, it does. In a real-effort experiment, Falk and Ichino (2006) find that average output is higher. Further, the standard deviation of output is lower within worker pairs than between pairs. Essentially, low-productivity workers raise their output towards that of their co-worker when a co-worker is present. One can imagine a number of possible explanations for this behavior, including subject uncertainty about the "true" compensation schedule. Similar results were found in a field experiment by Bandiera et al. (2009a), but *only* when the co-workers were friends. Specifically, in a situation where workers received individual piece rates and no appreciable production externalities existed, workers who were less able than a co-worker with whom they are friends increased their effort (and hence income) by 10%. In contrast to Falk and Ichino, however, Bandiera et al. also found that workers who are more able than their co-worker friends reduce their effort and forgo 10% of their earnings.[41]

A final, related question is how workers' effort changes when they can see each others' *wages*. Charness and Kuhn (2007) pose this question in a pure gift-exchange game in which workers knew that their productivity was different from their co-worker's, but did not know the size or direction of this difference. If between-worker equity concerns are important determinants of effort, we might expect that low-productivity workers (who tend to receive lower wages) would reduce their effort in treatments where they observe their co-worker's wage than when they do not (indeed this is suggested by Akerlof and Yellen (1990)). Perhaps surprisingly, they do not. Agents' primary concern seems to be to reciprocate generous wage offers from the firm; the authors speculate that responding to wage offers made to their co-workers would likely muddy this "signal" and is therefore avoided by workers.[42]

41 Mas and Moretti (2009) also detect small effects on productivity of being observed by a more productive worker, in a setting where workers were not paid for performance.

42 Maximiano et al. (2006) also examine gift exchange in multi-worker firms, but in their treatment all workers are equally productive and receive the same wage by design. Güth et al. (2001) consider a situation in which the principal can offer a menu of contracts to two independent agents with different productivities. As in Charness and Kuhn, public observability of the co-worker's contract induces the *employer* to compress compensation schemes.

In sum, laboratory tests of the one-period, one-agent, one-task principal-agent model have identified the following broad empirical regularities:

(a) Compensated wage cuts reduce effort in animal labor supply studies. In addition, leisure is normal, and uncompensated wage changes generate backward-bending labor supply curves (Battalio et al., 1981; Battalio and Kagel, 1985). For both animals and humans, there are large individual subject effects, both in the level of effort supplied for a given level of incentives, and in the *responsiveness* of effort to incentives.

(b) For humans, higher piece rates usually raise effort (Swenson, 1988; Sillamaa, 1999a,b; Dickinson, 1999). Agent self-selection into pay-for-performance schemes reinforces these effects (Lazear, 2000; Dohmen and Falk, 2006; Cadsby et al., 2007).

(c) Not paying at all can yield higher effort than low pay (Gneezy and Rustichini, 2000). Very high stakes can reduce agent performance in certain types of tasks (Ariely et al., 2008).

(d) Effort also responds to the *intercept* of the worker's compensation schedule, at least when generosity is seen as intentional (Fehr et al., 1993; Charness, 2004).

(e) Effort decisions can be affected by at least one type of reference point that can be manipulated in the lab: the subject's expected earnings for the session (Abeler et al., 2009).

(f) Forcible restrictions on agents' choice sets can reduce the efforts of agents on whom they are not binding (Falk and Kosfeld, 2006), but not when the restrictions are seen as "legitimate" (Schnedler and Vadovic, 2007).

(g) A decision by a principal to use piece rates can also reduce agents' efforts; a likely explanation is that the introduction of piece rates changes the agents' interpretation of the implicit contract for labor services (Irlenbusch and Sliwka, 2005a).

(h) A principal's decision to monitor the agent can also reduce agents' efforts, but only when the agents "know" the principal (Dickinson and Villeval, 2008).

(i) Unenforceable promises by principals to pay bonuses for "satisfactory" worker performance can elicit surprising amounts of effort, and can outperform more objective mechanisms such as random monitoring combined with punishment (Fehr et al., 2007b).

(j) Sharp discontinuities in reward schedules induce workers to misrepresent their output (Cadsby et al., 2008).

(k) Even when there is no strategic interaction between workers, workers' efforts may depend on their co-workers' efforts (Falk and Ichino, 2006). Co-workers' *wages* do not appear to affect effort, at least in our earlier work (Charness and Kuhn, 2007).

3.2. Tournaments

3.2.1. Theory

Consider now a situation where a firm employs multiple workers, who still do not interact in production. However, because the firm bases rewards, at least in part, on

agents' performance relative to each other, workers' effort decisions, pay levels, and utilities are interdependent. Relative performance plays a key role in a number of features of real-world compensation schemes, including promotions and bonuses for top-performing workers. Since at least Lazear and Rosen's seminal 1981 paper, economists have understood that, if workers are risk neutral, pay structures which award prizes based only on workers' relative performance can generate identical allocations in Nash equilibrium as would be achieved by optimal individual piece rates. This may be useful if rank order is easier to measure than cardinal performance, or if rewards are inherently indivisible (some promotions might be an example). Further, tournaments can be more efficient than piece rates if workers are risk averse and if their outputs are affected by a common shock. These results are robust to ability differences between workers if ability is public information: in that case, efficient tournaments typically include handicaps for the abler agents. As Carmichael (1983a,b), among others, has pointed out, payment by relative performance also mitigates an incentive problem affecting the principal, namely the incentive to understate workers' true outputs, or to provide suboptimal levels of complementary inputs after the contract has been signed. Finally, as O'Keefe et al. (1984) have pointed out, contests may have an efficiency advantage if workers derive direct utility from competition itself.

Counterbalancing the above advantages, tournaments may be less efficient than individual piece rates when workers' abilities are hidden information, especially if workers can self-select into tournaments (Lazear and Rosen, 1981). Also, in contrast to individual piece rates, tournament games in general require agents to think strategically about their co-workers' effort levels to find a Nash equilibrium; this may make them less robust as incentive schemes. It is also worth noting that, in general, optimal contest design will be different if (part of) the contest's objective is not simply to induce effort, but to identify the most talented contestant (e.g. for promotion). In this section we examine how tournament reward schemes work, not just theoretically, but when the games are played by human subjects in what has become a sizable experimental literature.

3.2.2. *Early experiments*

To our knowledge the first laboratory experiment on tournament-based incentives was by Bull et al. (1987). Bull et al. implemented tournaments between pairs of experimental subjects whose output was subject to independent, uniformly distributed productivity shocks (this guarantees a unique, pure Nash strategy equilibrium if the spread of the distribution is high enough), parameterized to yield identical equilibrium effort levels to a simple piece rate. As predicted, average effort levels were similar between the tournament and the piece rate, but effort variance across subject pairs was much greater under the tournament.[43] This suggests that, while equivalent in principle, tournaments may be a much less robust incentive scheme than piece rates in practice because agents

[43] This variance result appears quite robust. See, for example, Van Dijk et al. (2001) for a replication.

have difficulty finding a Nash equilibrium. Bull et al. also studied tournaments between players with different abilities, and found that less able agents systematically exerted more effort than the Nash equilibrium.

Schotter and Weigelt (1992) implement a very similar laboratory protocol but focus in more detail on "uneven" tournaments (where the participants' abilities differ), as well as on "unfair" tournaments (where the rules favor one identical agent over another). They are also interested in the effects of policies that (a) restore fairness in unfair tournaments (termed "equal opportunity laws"), or (b) give handicaps to less-able agents in uneven tournaments (termed "affirmative action laws"). Schotter and Weigelt find that (a) again, that mean effort levels in fair, symmetric tournaments match the Nash equilibrium; (b) disadvantaged contestants in unfair tournaments supply more than Nash equilibrium effort; (c) mean effort in uneven tournaments matched theoretical predictions, though largely because the less able agents either worked too much or chose zero effort, neither of which was Nash behavior; (d) symmetrizing previously-unfair tournaments raised both agents' effort levels; and (e) handicapping abler contestants raised total worker output and the principal's profit when the ability difference between contestants was large. It did so largely by eliminating drop-out behavior among the less able contestants. The authors suggest that these results might have some relevance to the effects of affirmative-action programs in the real world.

3.2.3. Selection into tournaments

The early theoretical literature on tournaments considered the question of selection into tournaments largely from an adverse-selection perspective. For example, Lazear and Rosen (1981) predicted that, if workers have private information about their own ability, less able workers will "contaminate" tournaments designed for abler workers. Whether or not firms modify the structure of their tournaments to address this adverse selection problem, the resulting effort allocations will no longer be as efficient as individual piece rates.

Experimental studies of self-selection into tournaments tend to focus on different questions. For example, building on earlier market-entry experiments in Industrial Organization (e.g. Rapaport, 1995), which tend to find surprising levels of co-ordination on the efficient outcome in entry decisions despite the absence of communication among subjects, Camerer and Lovallo (1999) designed a game in which MBA students chose whether to enter a "market" where their success depended on performance relative to other entrants, and on the subjects' own skill level (on a sample of logic puzzles or trivia questions about sports or current events). Camerer and Lovallo found excessive entry, which appear to stem not from inaccurate forecasts of the number of entrants or other factors, but from the subjects' substantial overestimates of their own ability. Vandegrift et al. (2007) conduct a similar experiment, where they allow a fixed population of agents to choose whether to be paid an individual piece rate or to enter a pool where they receive a prize of fixed value for the best performance. The task performed by workers

(forecasting the price of a fictitious stock based on cues that are correlated with the true price) is deliberately chosen to allow for "winning" to have some intrinsic or signaling value. These authors do not detect significant levels of excess entry. In this same vein, Niederle and Vesterlund (2007) also allow subjects to self-select between a piece rate and a tournament; their main interest is in gender differences. They find that, at given levels of ability, men exhibit significantly more overconfidence in their tournament entry decisions than women. We discuss this and related articles in more depth in the section on discrimination.

More recently, Eriksson et al. (2008) replicate Bull et al.'s (1987) original experimental design as closely as possible, with the exception that subjects are allowed to choose between a tournament (where they are randomly matched with another player) and an individual piece rate, calibrated to yield the same levels of optimal effort and expected utility. Thus, effort is a decision number and all agents are equally able. Eriksson et al., however, elicit subjects' risk aversion after the experiment. They find that risk-averse subjects are less likely to enter the tournament. This has the additional effect of reducing the high between-subject variance of tournaments, which was cited by Bull et al. as a possible disadvantage of tournaments. Mean effort was about one third higher in the tournament scheme; as in Lazear (2000) half of this was due not to incentive effects but due to selection. This is of particular interest here, since agents' abilities in this experiment were equal by construction.

3.2.4. Tournaments and risk-taking

Another prediction of tournament theory is that, in a tournament setting, agents have incentives to take actions that increase the spread of their output distribution, i.e. that increase risk. To our knowledge, Bronars (1986) was the first to discuss risk taking as a choice variable in tournaments; he argued that leading agents in sequential tournaments prefer a low risk strategy (to "lock in" their gains), whereas their opponents choose higher risk. Hvide (2002) shows that, even in a one-stage tournament, if there are no limits on risk taking, tournaments "collapse" in the sense that, for any given prize spread, agents choose infinite risk and zero effort in the Nash equilibrium. When there are limits on risk-taking, the Nash equilibrium is at the maximum level of risk; furthermore an exogenous reduction in the maximum permissible risk level raises both effort and welfare. Finally, Hvide shows that contests in which agents' performance is ranked according to its absolute distance from a target level k (thus outputs in excess of k are *punished*) have superior efficiency properties in this environment.

Agents' tournament-induced preferences for risk have been studied in various field contexts, including stock car racing—Becker and Huselid (1992) show that drivers take more risks when the prize spread is large—and investment fund managers (Brown et al., 1996; Chevalier and Ellison, 1997). Brown et al. find, as predicted by Bronars, that expected losers prefer high risks while expected winners prefer low risks. In the lab, Vandegrift and Brown (2003) find that high-variance strategies are indeed attractive in

tournaments, but primarily to agents with low capabilities performing a simple task. Nieken and Sliwka (2010) extend the theory of agent risk selection in tournaments to cases where the agents face correlated risks. They argue that—in contrast to Bronars' prediction—leading agents, rather than "playing it safe", may be forced to imitate their opponent's risky strategy. This prediction is confirmed experimentally. A possible application is to the case of mutual fund managers investing in the same, or similar, risky assets.

3.2.5. Sabotage

Since Lazear's important article (1989), economists have recognized that any compensation system based on relative worker performance rewards workers who take actions that reduce the measured performance of their peers, i.e. to engage in sabotage. Of course, sabotage is inherently difficult to study in the field, because workers may go to considerable lengths to conceal their acts of sabotage. This consideration has led a number of authors to study sabotage in the laboratory.

The first published laboratory experiment on sabotage appears to be Harbring and Irlenbusch (2005), who investigate sabotage in both a baseline treatment where the prizes are exogenously manipulated by the experimenter, and in a setting where principals in the experiment can choose the prize structure. Four agents compete against each other with the top two receiving a "winner" prize. Investing in sabotage reduces the output of all the other agents. When the prize spread is exogenously set by the experimenter, Harbring and Irlenbusch find, as predicted in simple tournament models, higher prize spreads encourage both greater effort and more sabotage; interestingly the latter effect dominates, suggesting that pay compression may be an optimal strategy. This finding does not, however, generalize to the case where prize spreads are selected by participants in the experiment, perhaps because in the authors' design higher spreads imply a higher expected value of the prize. Now, agents appear to reciprocate more generous compensation packages by refraining from sabotage. While it is not clear if this result would persist in designs that held the expected prize constant, the result does remind us that agents' "behavioral" intentions to reward or punish the *principal* may also play a role in real-world sabotage decisions.[44]

In a 2007 article, Harbring, Irlenbusch, Kräkel, and Selten consider sabotage in a contest where players are heterogeneous in ability. Three contestants play a two-stage "Tullock" contest where each agent selects *targeted* levels of sabotage aimed at each of the two other players. (Sabotaging other players makes it more costly for them to exert effort.) All sabotage levels are then revealed, and effort choices made in the second stage. In addition, players can be of two types—those with *ex ante* high effort costs ("favorites") or *ex ante* low effort costs ("underdogs"); this is publicly known in advance.

[44] Falk et al. (2008) also focus on the behavior of principals in a setting where agents can invest in sabotage. They find that both sabotage *and* loss-aversion among agents compromise the ability of large prize differentials to increase effort levels; principals respond to this by choosing wage compression.

Three treatments are implemented: one with homogeneous contestants, one with two underdogs and one favorite, and one with one underdog and two favorites. Contrary to expectations, when there are two underdogs, they do not "conspire" against favorites by directing their sabotage against the favorite. In line with expectations, an underdog engages in less sabotage when she is playing against two favorites than one; this resembles the dropout behavior of less-skilled agents in Schotter and Weigelt (1992). Harbring et al. also examine some treatments in which a saboteur's identity is revealed to the other contestants; they find that retaliation occurs in future rounds and that overall sabotage is less common.

Harbring and Irlenbusch (2008) implement tournaments with two, four, or eight contestants, and with the share of contestants who receive the winning prize equal to 1/4, 1/2, or 3/4. Agents choose both their own effort and a level of sabotage that affects all other agents equally. Parameters are chosen to yield identical Nash equilibrium effort levels in all treatments. Consistent with Nash behavior, neither tournament size nor the share of winning prizes has strong effects on effort, or on sabotage, though there is some tendency for effort to be higher when the share of winning prizes is 1/2, compared to 1/4 or 3/4. In a more recent paper (2009) the same authors introduce communication among the principal and agents. This turns out to curb sabotage via agreements on flat prize structures and increased output.

To our knowledge, the only papers to study sabotage in a real-effort experiment are Carpenter et al. (2010) and Charness et al. (2010). Sabotage in Carpenter et al.'s context takes the form of peers' subjective evaluation of the quality of each others' output. (The task is printing letters, placing them into envelopes, and handwriting addresses on envelopes. Subjects also had an opportunity to miscount the number of envelopes produced by their co-workers.) Carpenter et al. found that subjects responded to the possibility of sabotage by their co-workers by producing less output than when peer review was not possible. Piece rate compensation performed much better than the tournament with peer-review. In Charness et al. (2010), people are paid a flat rate for their work (the task consists of decoding sets of one-digit numbers into letters from a grid of letters that is displayed on the computer screen), and learn about the rank of their production in a 3-person group. In one treatment, people could pay to sabotage the production of the other people in the 3-person group and many people did so.

3.2.6. Collusion

In all tournaments, agents have an incentive to collude against the principal: if they can all agree to exert zero effort, the prize in any fair tournament will be randomly assigned to one of them, and they will all be better off than if they had truly "tried" to win the prize. Despite this feature, collusion rarely appears to occur in tournament experiments.[45] This

[45] Of course, while zero effort by all agents is probably a good sign of collusion, the question of how one would identify collusive behavior in a tournament is an interesting one. Clearly, individual effort can be below the privately-optimal level for reasons other than collusion.

may be because various features of the design, including anonymity and re-matching, are deliberately chosen to make collusion difficult. To our knowledge, Harbring and Irlenbusch (2003) is the only laboratory experiment that addresses the issue of collusion. They find (as one might expect) that the smallest (two-person) tournaments are the most conducive to collusion. It would be interesting to see if greater amounts of collusion are observed when there is less anonymity and more opportunity for communication and repeated interaction among agents.[46]

3.2.7. Feedback

Suppose that the production process takes time during which partial information about the agents' relative performance becomes available. How will this information affect agents' subsequent effort decisions? Theoretically, even with standard preferences, this depends on both the distribution of the randomness in the effort-output relationship and the shape of agents' disutility-of-effort functions; thus Eriksson et al.'s (2008) recent finding that feedback has no effect is not necessarily surprising.[47] More recently, however, Gill and Prowse (2009) consider a case—where the probability of winning is linear in the difference in the agents' efforts—where such information should have no effect under standard preferences; despite this, they find that agents who are behind exert less effort.[48] They interpret this "discouragement effect" as a consequence of disappointment aversion, and estimate both the level and heterogeneity of disappointment aversion in their sample using structural methods.

Using field data from over 60,000 basketball games, however, Berger and Pope (2009) find the opposite: being slightly behind at halftime leads to a discontinuous *increase* in a team's winning percentage; this apparent psychological effect is roughly half the size of the home-team advantage. This field data is corroborated with experimental evidence; the experiment sheds important additional light on the field data by helping to rule out alternative explanations. The contrasting results of these three studies on feedback suggest there is much we still do not understand about the effects of providing interim performance feedback on effort decisions.[49]

[46] In an interesting field experiment, Bandiera et al. (2005) found that, when engaged in a tournament with their friends, fruit pickers moderated their output (relative to a piece rate) in apparent response to the negative externality their effort imposed on those co-workers. Since this behavior occurs only when workers can monitor others and be monitored, it seems more likely to be motivated by collusion than altruism.

[47] Eriksson et al.'s finding that feedback tends to reduce the *quality* of the low-performers' work is however evocative of the greater risks taken by agents who find themselves running behind, discussed earlier in this section.

[48] In Gill and Prowse's context (where the agents select their outputs sequentially, and the interim feedback consists of informing the second agent about the first agent's performance), being "behind" simply corresponds to a high level of performance by the first agent.

[49] An additional question concerning interim feedback involves the *principal's ex post* incentives to reveal this information honestly; for example, it may be in the principal's interest to report that the race is closer than it really is. Both Gürtler and Harbring (2007) and Ederer and Fehr (2007) consider this question; they do find evidence that the gap is underreported, though the level of underreporting is lower than predicted by a model without aversion to lying.

In sum, laboratory tests of the tournament models have identified the following empirical regularities:

(a) A properly-designed tournament can replicate the results of an efficient piece rate in expectation, but generally yields greater variance in mean output across agent groups (Bull et al., 1987).

(b) Handicaps, or "affirmative action" tend to improve the performance of tournaments between unequal agents. Part of this effect, however, is due to non-Nash choices by less able agents in the absence of handicaps (sometimes working too hard, sometimes dropping out completely) (Schotter and Weigelt, 1992).

(c) Under some conditions, decisions to enter into tournaments are surprisingly close to optimal levels (Rapaport, 1995). For certain populations, however, entry can be excessive due in part to overconfidence. See Camerer and Lovallo (1999) for MBAs and Niederle and Vesterlund (2007) for men. Allowing risk-averse agents to self-select out of tournaments reduces the between-group variance in output (Eriksson et al., 2008).

(d) Tournaments can increase risk-taking (Vandegrift and Brown, 2003); this effect is not necessarily confined to agents with a low probability of winning (Nieken and Sliwka, 2010).

(e) Increases in tournament prize spreads can raise sabotage as well as effort; this effect can be strong enough to reduce total output (Harbring and Irlenbusch, 2005; Carpenter et al., 2010). Agents do not always "target" their sabotage in the expected direction (Harbring et al., 2007).

(f) Collusion is rare in anonymous tournaments with more than two contestants (Harbring and Irlenbusch, 2008).

(g) The effects of interim performance information on agents' subsequent efforts remain poorly understood. Recent studies have found intriguing effects in opposite directions, with Gill and Prowse (2009) finding that being behind reduces effort (when standard preferences predict no effect) and Berger and Pope (2009) finding the opposite.

We conclude this section by noting one aspect of tournaments that has so far received relatively little attention in the lab. This is the use and effectiveness of tournaments as a tool, not for eliciting efficient amounts of effort, but for *identifying the more able player*. (A recent exception in the field is Calsamiglia et al., 2009). If—as Gibbons and Waldman (1999) argue—it is efficient for organizations to promote abler persons into higher-level positions, then the almost-exclusive focus of the experimental literature on tournaments' consequences for effort levels (as opposed to efficient inference of the agents' underlying abilities) may be missing a key function of relative performance evaluation schemes in real organizations.

3.3. Teams

3.3.1. Holmstrom's model

In the classic model of agency in teams (Holmstrom, 1982), a group's output, x, is a differentiable function of the effort levels of its n members, e_i. The principal's problem is to design a set of compensation functions, $w_i(x)$, that depend only on the group's total output and induce efficient effort choices by *all* agents. Holmstrom's well known result is that such a function cannot exist if $w_i(x)$ takes the form of a "sharing rule". A group compensation function is a sharing rule if it satisfies $\sum_{i=1}^{n} w_i(x) = x, \forall x$, i.e. it balances the budget for all possible group output levels, not just the group's equilibrium output. Thus, free riding is inevitable unless the principal can commit to paying agents as a group more than their combined output for some out-of-equilibrium effort choices, and less in others.

To our knowledge, the first economists to study team production in the lab were Nalbantian and Schotter (1997).[50] In their experiments, Nalbantian and Schotter create two teams of six workers each whose group output is given by the formula $x = \sum_{i=1}^{6} x_i + \varepsilon$, where ε is a uniformly distributed random variable. Four compensation schemes are compared: revenue sharing (where each worker is paid $1/6$ of the group's output), "forcing" contracts or targets (where revenue is shared *if* group output exceeds a target; otherwise all agents receive a low payoff), "gainsharing" (where the target is a function of the team's past performance in the experiment), and a tournament between the firm's two teams (where all members of the team with the higher output receive a prize). Consistent with the "classic" model's predictions, effort levels under the revenue sharing scheme converge towards individually rational levels, which entail a high degree of free riding (though they start out considerably higher in early rounds). Forcing contracts (including gainsharing) perform poorly, perhaps because of the multiple equilibria that are theoretically associated with the induced game among agents. Competition between teams generates the highest level of efficiency among all these treatments, at least for Nalbantian and Schotter's parameterization, which sets a prize high enough to (theoretically) induce efficient effort levels.[51]

More recently, Meidinger et al. (2003) have introduced "active" principals into Nalbantian and Schotter's protocol: here, a subject in the role of the principal first makes an offer to two prospective team members; the offer is the share of the team's output that the team gets to keep—either one half or two thirds. Agents then decide independently on effort levels. In addition to the usual cooperation versus free-riding considerations,

[50] The team production problem is, however, closely related to the problem of voluntary contributions to a public good, which has a longer experimental history.

[51] Nalbantian and Schotter also implement a scheme where the principal can observe an *individual's* effort at a cost. Not surprisingly, they find that "monitoring works, but is costly". Van Dijk et al. (2001) implement individual, team, *and* tournament based compensation in a real-effort experiment. While some free riding occurred in teams, overall effort levels were the same under individual and team payment, since free-riding was counteracted by many subjects providing more effort than in case of individual pay.

agents' effort decisions in this context also appear to be affected by a desire to reciprocate generous "share" offers from the principal.

3.3.2. Team production and the voluntary contributions mechanism (VCM)

A second branch of the "team production" literature traces its origins not to the principal-agent literature but to that on public goods. Team production in these studies is defined as contributors' behavior in a voluntary contributions mechanism (VCM) (see Isaac and Walker, 1988 for an early example; Ledyard, 1995 for a review of the experimental literature, and Chaudhuri, 2007 for a more recent review of specific issues). In the standard VCM, each member of a group, i, voluntarily contributes m_i "tokens" out of his endowment, w_i to a common account. Each member's payoff is then just $(w_i - m_i) + a\sum_i m_i/N$, where N is group size and a is the efficiency gain from public provision. Thus, in the "classic" VCM, the reward schedule for the agents is predetermined to be a sharing rule, with equal shares accruing to all agents. The typical experimental result for the basic VCM is that individual contributions start out above the individually-rational (but socially inefficient) level but converge to that level as agents gain experience with the game. Work within this tradition has examined the impact of factors like group size, communication, and group heterogeneity on voluntary contribution levels; see Chaudhuri et al. (2006) for a recent example. Recent summaries of results in the VCM literature are available in Plott and Smith (2008, Chapters 82-90).

Beginning in the 1980s, some experimenters interested in team production in the workplace studied effort decisions using a VCM framework. The key modification to the VCM framework that was introduced by these researchers is reward schedules that do depend in some way on the effort contributions of individual team members.[52] Since we know (at least theoretically) that either individual piece rates or tournaments can induce efficient effort levels in this case, the main questions concern the efficacy of particular types of reward schedules in raising individual contributions toward efficient levels. Examples of this approach include Dickinson and Issac (1998), who introduce prizes for the highest individual contribution into a standard VCM environment; not surprisingly, these prizes raise contribution levels. (Of course, to be effective, such prizes must be committed to regardless of the level of output that is attained by the group; thus they constitute precisely the kind of "budget-breaking" that Holmstrom (1982) showed is necessary to attain efficiency in a team production environment.) Other papers who have introduced different types of relative rewards into VCMs are Dickinson (2001), which considers monetary *fines* on the lowest contributors, and Irlenbusch and Ruchala (2008).

[52] Eckel and Grossman (2005) consider a classic VCM problem, framed as a team production problem, with no modifications to the incentive structure at all. Instead, they manipulate the amount of anonymity, contact and "identity" of team members, for example by having the members perform a cooperative task before the experiment. They find some positive effects of increased contact among team members, though the two interventions that yielded the greatest gains also offered extra financial incentives relative to the base case (wages for teamwork, and a monetary bonus to the team with the highest total output).

3.3.3. Complementarities in production

An ironic feature of most models of team production that have been implemented in the lab (including the basic VCM) is that the assumed production technology—in which all agents' efforts are perfect substitutes for each other—rules out one of the main reasons why teams exist: production complementarities among the members. An extreme form of complementarity that has, however, been studied in the lab is the "minimal effort" or "weakest link" game (Van Huyck et al., 1990) where group output depends only on the lowest effort supplied among the team members. Especially in the absence of communication among agents, deductive methods provide little guidance regarding what equilibria to expect in such co-ordination games; Van Huyck et al. find massive co-ordination failures in the lab as agents play "safe" strategies that leave them relatively invulnerable to the hard-to-predict actions of their co-workers.[53]

More recently, Brandts and Cooper (2007) consider Leontief production in a team of four workers. With no communication or management, Brandts and Cooper show that—as in earlier studies—such games almost always converge on co-ordination failure, with all agents supplying low effort levels. The authors then assign a manager to each such team, and give the manager two types of tools: (a) increasing the (common) rate of pay received by each team member for an extra unit of team output—recall that both the manager and the workers see only the team's total output, which equals the lowest effort level chosen by any team member,—or (b) communicating with the workers by sending (and in one treatment also receiving) completely unstructured messages to the group. All of this communication takes place before any effort decisions are made.

Brandts and Cooper's provocative finding in this environment is that increased financial incentives are essentially powerless in raising group performance (small increases in group output can be achieved, but never enough to compensate the principal for the cost of the additional compensation). Unenforceable messages, however, can be highly effective in overcoming co-ordination failures of the type modeled by these authors. The authors supplement their experimental analysis with an interesting econometric analysis of the effects of different types of messages exchanged between principals and agents.

Most recently in this context, Georg et al. (2009) test an intriguing theoretical result about incentives in teams with production complementarities due to Winter (2004). Winter's result refers to a group of n identical agents contributing either 0 or 1 unit of effort to a project, each of whom receives a reward of b_i, $i = 1, \ldots, n$ if the project succeeds and 0 otherwise. The production function is said to exhibit complementarity when $p(k+1) - p(k)$ increases in k, where k is the number of agents exerting positive

[53] Bornstein et al. (2002) find that this co-ordination is somewhat mitigated when intergroup competition is introduced—specifically, two groups compete for a prize received by the group with higher minimum. The members of the losing group were paid nothing. Sutter and Strassmair (2009) introduce communication into experimental tournaments between teams. In their experiment, communication within teams increases efforts (by facilitating coordination) while communication between teams reduces effort (by facilitating collusion).

effort and p is the probability of success. In this situation, Winter shows that the *only* reward scheme that induces efficient effort as a unique Nash equilibrium when workers' efforts are complementary is fully discriminating, in the sense that no two workers' rewards for group success, b_i, are the same. Thus, efficiency *requires* inequity, in the sense of treating identical workers differently. Intuitively, it is easier (and cheaper) to make some agents work if they "know" that working is a dominant strategy for some other (identical) agent because that agent will be extremely well rewarded if the entire group succeeds.[54]

Of course, given many authors' arguments that workers' effort decisions can be strongly, and negatively, influenced by 'unfair' wage differentials (e.g. Akerlof and Yellen, 1990), it is not at all obvious that Winter's proposed mechanism—which *requires* arbitrary wage differentials—would work well in practice. Perhaps surprisingly, it does: When the production technology exhibits complementarity, higher efficiency is achieved under a discriminatory reward mechanism than under a cost-equivalent symmetric one. Further, despite concerns that fairness considerations might affect workers' behavior, "subjects' effort choices are highly sensitive to their own reward, but largely unresponsive to the rewards of the other . . . subjects in their group". This echoes Charness and Kuhn's (2007) result for individual labor contracts, which also found that workers' effort decisions in a multi-worker firm were insensitive to other workers' wages. Taken together, these two papers suggest that, in contrast to the role of social preferences in exchanges between individual "workers" and "firms", the importance of horizontal comparisons (*among* workers) within firms seems limited in the experimental literature.

3.3.4. Selection into teams

As noted, probably the most common team compensation policy is a sharing rule in which all members receive an equal share of the group's output. An immediate consequence of such a policy—which contrasts starkly with tournament-based compensation—is that team members will prefer to have *abler* co-workers. Thus the issue of how teams are formed can be an interesting one, especially when complementarities exist among team members: what are the effects of different team formation mechanisms on the matches that are formed, and are these matches efficient?

We are aware of only a handful of laboratory studies of the team-formation process. Weber (2006) conducts experiments on the minimum-effort co-ordination game that start with small groups—who find it easier to co-ordinate—then adds entrants who are aware of the group's history. Using this procedure, coordinated large groups can be created, as long as the rate of growth is not too large. In contrast to Brandts and Cooper (2007), no communication is involved. Charness and Yang (2008) evaluate a specific voting mechanism for group formation in a VCM where there are economies of

[54] In the opposite case of substitutability in production, the only reward scheme that induces efficiency as a unique Nash equilibrium is the symmetric one that treats all workers alike. Thus, Winter's model provides interesting predictions for the relationship between the group's production technology and the efficient compensation policy.

scale. Societies of nine people are initially formed randomly into three groups of three people who play the game for three periods. Individuals then learn about the average contribution of each individual (by ID number) in their current own group, as well as the average contribution in other groups, and can decide whether to exit the group. Remaining group members choose whether to exclude any current members from the group; the new groups and "free agents" then choose whether to merge with other existing groups and/or other free agents. A critical element is the role of efficiency in terms of group size—the multiplying factor for each token contributed to the group account increases with group size. Groups of nine are common, particularly after a 'restart' after 15 periods. They find considerable success for the mechanism, as the threat of ostracism seems to keep contribution rates quite high and efficiency is a driving force.

In an interesting field experiment, Bandiera et al. (2009b) find that, when the incentives facing an entire team are strengthened (in their case by introducing a tournament between teams or performance feedback), assortative matching into teams by ability is increased. Further, workers become less likely to form teams with those they are socially connected to. In Hamilton et al.'s (2003) field study of a textile plant, strong assortative matching into teams did not occur. Further, equal sharing of production bonuses within teams seems to have stimulated cooperation, information sharing, monitoring and even mutual training within teams, generating a productivity increase (relative to piece rates) despite the expected free-rider problem.

Finally, in addition to complementarity between team members' effort levels, there may also exist complementarity of a different sort—actual gains from heterogeneity, for example if different types of group members possess complementary skills. Charness and Villeval (2009) cite some work that suggests this and also finds that there is a preference for mixed ages (and higher efficiency) in a VCM played at two French firms' work sites.

In sum, laboratory tests of effort decisions in teams have identified the following stylized facts:

(a) In the absence of communication and/or repeated interaction, teams in which agents are paid equal shares of the team's output perform poorly, with agents' efforts converging to low, individually rational levels after a few rounds of play (Isaac and Walker, 1988; Nalbantian and Schotter, 1997).

(b) The forcing contracts (essentially group bonuses) suggested by Holmstrom (1982) typically fail to improve outcomes in these environments due to co-ordination problems among agents (Nalbantian and Schotter, 1997).

(c) Team performance may also be affected by considerations of reciprocity towards the principal, if one exists (Meidinger et al., 2003).

(d) Adding incentives based on the relative contributions of individual members to the team's output can improve teams' performance, if such measures are available (Dickinson and Issac, 1998; Dickinson, 2001; Irlenbusch and Ruchala, 2008).

(e) Adding competition between teams can be more effective than any of the above strategies (Nalbantian and Schotter, 1997). Given the tremendous popularity of team sports, both to participants and spectators, it is not at all implausible to us that humans are naturally attracted to such situations and perform well in them.

(f) When there is complementarity between the efforts of team members, loss of output due to co-ordination failures can be severe in the absence of communication among team members (Van Huyck et al., 1990). Adding "cheap talk" communication in such situations can generate dramatic improvements, much more so than strengthening financial incentives (Brandts and Cooper, 2007). Other mechanisms that have been observed to work include asymmetric incentives—that facilitate co-ordination by making high effort a dominant strategy for at least one player— (Georg et al., 2009), and slowly adding new members to smaller groups, which find it easier to co-ordinate (Weber, 2006).

We conclude by noting one aspect of team production that, to our knowledge, has not been addressed in the lab: opportunistic behavior by principals. For example, principals who attempt to commit to incentives that "break the budget" at out-of-equilibrium effort levels may face strong temptations to understate the team's total output (see for example Eswaran and Kotwal, 1984).[55]

3.4. Multi-task principal-agent problems

Suppose the agent performs multiple tasks the principal cares about, but the principal is only able to base the agent's compensation on her performance in a subset of those tasks. At least since Farrell and Shapiro (1989) outlined their "Principle of Negative Protection", and Holmstrom and Milgrom (1991) introduced the "multi-task principal agent problem", economists have recognized that, depending on circumstances, incentive systems based on a subset of the tasks performed by the agent may be less efficient than a compensation system with no incentives at all.

The first experimental implementation of a multi-task principal-agent model we are aware of is Fehr and Schmidt (2004).[56] In their experiment, an agent exerts two types of effort, both of which are observed by the principal and agent, but only one of these is contractible. The two types of effort are complements in the production of total output (specifically, output equals the product of e_1 and e_2). Disutility of effort, on the other hand, is given by $c(e_1 + e_2)$, where c is increasing and convex. Principals can choose between two types of contracts: A "piece rate contract" pays the agent a fixed base wage plus a linear piece rate per unit of task one performed. A "bonus contract" consists of

[55] Note that tournaments do not face this problem, since the principal's total compensation bill does not depend on which worker wins; indeed this has been suggested as a possible advantage of tournament-based compensation (Carmichael, 1983a,b).

[56] There is, of course, a sense in which tournaments in which agents can allocate their effort between productive activities and sabotage are a multi-task situation. We discuss these in the section on tournaments.

a fixed base wage plus an unenforceable announcement that the agent might receive a bonus from the principal if his overall performance is "satisfactory". Principals and agents interact anonymously, and only once.

Absent concerns for reciprocity (in the sense of perceived intentions for offering a bonus rather than a fine), it is clear that both the above contracts should perform relatively poorly: the bonus contract should yield zero effort, and the piece rate should produce effort only on the first task, which leads to very low output due to the assumed complementarity in production. Allowing for reciprocity, it is not immediately clear which contract should perform better, since both contracts allow for principals to make generous fixed wage payments that agents could conceivably reciprocate. Empirically, however, echoing results in Fehr et al. (2007b), bonus contracts perform much better than piece rates in this environment: many agents reward high effort levels on both tasks with generous bonuses, and agents seem to anticipate this. (Some apparent reciprocation of high fixed pay was evident in both treatments, but was not very effective, relative to the possibility of earning a bonus.)

In some sense, Fehr and Schmidt's results support Holmstrom and Milgrom's prediction that powerful incentives may be a mistake when they are based on a strict subset of the agents' actions that affect the principal's welfare. Indeed in such situations, Fehr and Schmidt's results suggest that vague and unenforceable subjective performance evaluations by the principal can outperform piece rates in this case.[57] Of course, the advantages of such "bonus" contracts depend heavily on what Fehr and Schmidt interpret as workers' concerns for fairness vis-à-vis the firm; if workers' perceptions of fairness are manipulable or highly context-dependent, the widespread use of vague expectations as a solution to multitask principal-agent problems may not be practical.

The only other multitask principal-agent experiment we know of is by Oosterbeek et al. (2006). In contrast to Fehr and Schmidt, rather than choosing between two complementary, productive activities, Oosterbeek, Sloof and Sonnemans' agent chooses between a productive and a "rent-seeking" activity. The latter activity is pure social waste, but increases the agent's bargaining power by improving his outside options. In the "classic" model with no social preferences, effort devoted to rent seeking should increase when the experimenters raise the marginal efficacy of effort in that activity. In some behavioral models, however, the opposite could happen, because abstaining from rent-seeking activities becomes a more powerful signal of the agent's good intentions the more effective those activities are. The authors' experimental results are largely in line with the classic model.

In sum, economists are just beginning to study the multitask principal-agent problem in the lab.[58] Some key patterns that have been observed to date are:

[57] See for example Baker et al. (1994) for a more explicit argument in favor of subjective performance evaluation.

[58] Field and econometric studies are also rare. Slade (1996) studies multitasking in contracts between oil companies and gasoline stations; more recently Griffith and Neely (2009) consider multitask incentives in a UK distribution firm.

(a) As predicted by the "standard" model (Farrell and Shapiro, 1989; Holmstrom and Milgrom, 1991), rewarding the observable task via a piece rate while not rewarding the other yields poor outcomes, especially if the tasks are complements (Fehr and Schmidt, 2004).

(b) Unenforceable promises by the principal to reward "satisfactory" *overall* performance by the agent perform remarkably well (Fehr and Schmidt, 2004). Perhaps this situation is familiar to subjects and they act according to norms that are highly effective in the real world.

(c) When agents can choose to invest in an unproductive rent-diverting activity, raising the returns to that activity generates more of it. Agents seem neither to anticipate, nor to receive, increased rewards for refraining from such activity when it becomes more tempting for them to undertake it (Oosterbeek et al., 2006).

3.5. Multi-period principal-agent interactions

3.5.1. Ratchet effects

The "ratchet effect" applies to a situation where a principal contracts with an agent more than once, the agent has some persistent private information (such as his ability or the productivity of the principal's technology), and binding multi-period contracts are not enforceable. In such situations, actions taken by the agent early in the relationship reveal information to the principal, which can be used by the principal later on to the agent's disadvantage. The classic example in labor economics is in the context of piece rates, where an agent's choice of a high effort level in the first period reveals either that (a) the agent's effort costs are low, or (b) the firm's technology is more productive than expected, either of which leads the firm to reduce the generosity of the agent's compensation package in the future. Anticipating this, "able" workers (or workers who have discovered that the firm's new technology is highly productive) will choose low first-period effort levels (Gibbons, 1987; Ickes and Samuelson, 1987). This benefits those workers by preventing the firm from extracting their rents later in the relationship, but is socially inefficient.

Aside from some early ethnographic studies (see for example Mathewson, 1931), the *only* empirical evidence of ratchet effects of which we are aware is experimental in nature.[59] Chaudhuri (1998) conducted a laboratory experiment in which principals and agents interacted for two periods, and agents were one of two types that were unobserved by the principal. There was little evidence of ratcheting: most agents played naively, revealing their type in the first period even when an informed principal would use this information to the agent's disadvantage, and principals often did not exploit agents' type revelation. Possible explanations for this result include the relative complexity of the game, and the lack of context provided to the subjects that might have impeded the learning process.

[59] Carmichael and MacLeod (2000) interpret some historical evidence in light of the ratchet effect.

Cooper et al. (1999) frame their experiment in a context-rich way, as a game between central planners and firm managers, use both students and actual Chinese firm managers as subjects, and implement experimental payoffs with high stakes relative to the participants' real-world incomes. They also simplify the interactions between principals and agents, focusing the experiment only on the stages of the game where information revelation matters: the agent's effort choice in the first period, and the principal's choice of a payoff schedule in the second. Cooper et al. do find evidence of ratchet effects, though even in their context it took some time for the players to learn the consequences of type revelation.

Finally, Charness et al. (2008) experimentally test a prediction of Kanemoto and MacLeod (1992) that *ex post* competition for agents can eliminate ratchet effects and lead to first-best outcomes in equilibrium. Importantly, this prediction holds even when outside firms cannot observe the past performance of the agent. They also extend Kanemoto and MacLeod's theoretical analysis to show that *ex post* competition for principals has the same effect. They impose three conditions in their experiment: no *ex post* competition, competition with an excess supply of principals, or with an excess supply of agents. As predicted, both types of competition virtually eliminate ratchet effects, though of course their effects on the utilities of the agent differ dramatically.

3.5.2. Career concerns

In ratchet effects models, high-ability agents exert low effort in early interactions to convince principals they have low ability; this prevents principals from reducing the agents' compensation in the future. In "career concerns" models, such as Holmstrom (1982, 1999) agents exert high effort in early interactions, in order to convince principals they have high ability, and thereby attract high wage offers in the future. The key reasons for these dramatic differences are different assumptions about information and competition: In ratchet models, the principal and agent are in a bilateral monopoly situation, and the agent's performance is typically not seen by other firms; thus high effort by the agent signals that he has a high level of rents that can be extracted in the future. In career concerns models, there is a competitive market for agents in all periods and the agent's output is seen by other firms; thus the wage the agent can command in future periods increases with the market's assessment of his ability. In the literature, ratchet models have been used to describe long-term worker-firm interactions, or interactions between managers and planners in non-market economies. Career-concerns models have been applied mostly to the compensation of CEOs and other senior executives, beginning with Fama (1980).

We are aware of two experimental papers on career concerns; the more recent (Koch et al., 2009) is actually the simpler one. Consider a world with a finite number of agents and principals. In the first period, no agent's ability is known to anyone, and all principals and agents share the same prior for agents' abilities. In this period, an agent chooses an effort level e at a private cost, $c(e)$, which is common to all agents. Once effort has been

chosen, the worker's output is publicly revealed to be $y = a + e$, where a is realized worker ability. A worker's ability is a permanent characteristic, thus in this simple version of the model a worker completely learns his ability at the end of the first period. At the end of period 1, all firms see every worker's first-period performance y, and will rationally attempt to infer each worker's ability from this information.

In period two, agents exert no effort but produce an output equal to their ability, a. Knowing this, principals engage in (Bertrand) competition for agents at the start of this period. (In the experiment, four firms bid for three agents; a principal can employ more than one agent; principals simply offer lump-sum wages that agents accept or not.) In a Perfect Bayesian equilibrium, firms will therefore offer wages equal to agents' abilities in the second period; this gives agents incentives to take period-one actions that convince firms they have higher abilities. Subject to certain restrictions on out-of-equilibrium beliefs, Koch et al identify a unique Perfect Bayesian equilibrium in which all agents choose the same, strictly positive effort level in period one. In consequence, first-period outputs perfectly reveal agents' abilities, and agents are paid their abilities in period two. While all agents would be better off choosing zero effort in period one (and principals earn zero profits in all equilibria), the logic of "signal-jamming" equilibria such as these compels all agents to work harder, to prevent being misidentified as a lower-ability type.

In their experiment, Koch et al. compare the above model to a "public ability" treatment, which is identical except that both the worker's output and effort (and therefore ability) are publicly revealed at the end of the first period. In this treatment agents have no reason to expend effort in the first period, so equilibrium effort is predicted to be zero. By and large, Koch et al.'s findings are consistent with the predictions of the career concerns model: effort is higher in the hidden-ability treatment, and subjects' first and second-order beliefs (which are elicited by the experimenters) are quite consistent with the model. That said, decision errors were high in early rounds (it apparently took some time for subjects to understand the game), and principals' offers were subject to a mild winner's curse.

Irlenbusch and Sliwka's (2006) model of career concerns is identical to Koch et al.'s, except that it contains two key features. First, firms bid for workers at the start of the first period as well as the second. Second, agents choose effort in the second period as well as the first. Since all workers are *ex ante* identical (so there is no reason to offer them different wages in the first period), and since all workers' privately optimal second-period effort is zero, neither of these differences changes the PBE of the game. Interestingly, however, the experimental outcome is dramatically different: now, contrary to predictions, first-period effort is much higher in the public ability treatment, precisely where there is no signal-jamming reason to exert effort. According to both Irlenbusch and Sliwka and Koch et al., the most likely reason is that Irlenbusch and Sliwka's model introduces opportunities for signaling of a different kind (that is not formally modeled in either paper): because effort is now revealed before period-two wage offers are made, agents

now have an opportunity to signal that they are "high-effort" (or "reciprocal", or "fair") types by choosing high effort in the first period. Since effort is not directly observed in the hidden-ability treatment, such signaling is less effective there. It is noteworthy, however, that this apparent signaling behavior occurs even though, by construction, all agents have the same cost of effort function in the experiment. The effort costs, or "willingness to work" that agents are apparently signaling, is some personal characteristic that is not induced by the experimental design. (Though it is clearly consistent with heterogeneous social preferences.)

Of course, which of the two above designs is more representative of any particular "real world" labor market is unclear. While it is clear that bidding for workers and effort decisions are made at multiple points in any worker's career, Koch et al. argue that, to the extent that agents' actions in Irlenbusch and Sliwka's experiment are driven by reciprocity or fairness concerns, they may be irrelevant to the market for CEOs. On the other hand, if these actions are meant to signal a low cost of effort, or a high level of determination, drive and ambition, they may indeed be very relevant to the CEO market.[60]

3.5.3. Investments and hold-up

The "hold-up" problem pertains to any multi-period relationship involving specific investments, where binding multi-period contracts are not enforceable. An early formal statement was in the context of a unionized firm's investments in plant and equipment, where market frictions create a gap between capital's purchase price and its resale value (Grout, 1984). Grout showed that the firm in general under-invests when union-firm bargaining occurs after the capital is in place, since the firm pays the full cost of the investment but (due to *ex post* surplus sharing) it only reaps a fraction of the returns.

Most applications of the hold-up problem in labor economics, however, refer to the problem of *workers'* investments in firm-specific skills. Long ago, Becker (1964) proposed that sharing both the costs and returns to firm-specific investments should achieve (presumably constrained) efficiency in separation decisions after the investment is made, though the precise efficiency properties of this arrangement were not specified, nor were its implications for the initial investment decision analyzed. Since then, Hashimoto and Yu (1980), Hall and Lazear (1984) and others have studied the problem more formally, and some ingenious institutional solutions (such as rigid wages in combination with a "triggered" renegotiation (MacLeod and Malcomson, 1993) have been proposed. Other institutional arrangements that have been argued to help solve the firm-specific

[60] Akerlof (1976) presents a signaling model in which agents signal a low cost of effort by working harder; separating equilibria in which (almost) all workers provide socially excessive effort exist. For a more recent career concerns model with similar properties, see Acemoglu et al. (2008). It may also be worth noting that labor economists have developed a variety of models in which workers' abilities are gradually revealed to markets over time (see for example Bernhardt and Scoones, 1993, and Altonji and Pierret, 2001). Since our focus in this section is on principal-agent interactions, we consider only the subset of those models which, like the career-concerns model, involve an effort decision by workers.

training problem (in the sense of inducing both efficient investments and efficient *ex post* separations) include various types of promotion ladders (Carmichael, 1983a,b; Prendergast, 1993), and multi-skilling policies (Carmichael and MacLeod, 1993).[61] We note that a key feature of all these analyses is that principals (firms) cannot make binding commitments, for example to retain a worker if that is *ex pos*t unprofitable for the firm. If firms can make credible promises of this nature (perhaps because their interactions with previous cohorts of workers can be observed by current workers), the holdup problem can be substantially mitigated, or eliminated.

The first experimental study of the hold-up problem of which we are aware is Anderhub et al. (2003). They consider a two-period worker-firm interaction in which workers choose whether to make a firm-specific investment that reduces their effort costs in the first period; the investment is profitable to the worker only if he is employed by the same principal in both periods. In perfect equilibrium, workers should make the investment whether or not firms can commit to re-employ them; perhaps unsurprisingly, Anderhub et al. however find that contractual form does matter, with workers being more willing to make specific investments when their re-employment is contractually guaranteed. More recent experimental papers have examined the effects on holdup of communication (Ellingsen and Johannesson, 2004); of which party (worker or firm) makes the investment and the nature of *ex post* wage bargaining (Oosterbeek et al., 2007a); and of firms' promotion rules (Oosterbeek et al., 2007b).

More specifically, the promotion rules compared in the latter paper are "up or stay" (Prendergast, 1993), versus "up or out" (Kahn and Huberman, 1988); the setting is one in which workers' specific investments may affect their firm-specific productivity not only in their current job, but in an alternative job (in the same firm) to which they might be promoted. Promotion decisions are at the firm's discretion, and firms cannot commit to refrain from opportunism in those decisions (aside from being bound by an up-or-out or an up-or-stay rule). Workers' *only* incentives to acquire firm-specific skills in these models are to win promotions or to keep their jobs. In this context, up-or-stay promotion policies will induce workers to invest only if investments raise their productivity more in the job to which they would be promoted than in the current job. Thus investment incentives may be too weak, but specific investments are never wasted due to separations. Up-or-out policies can provide better investment incentives, but may waste investments. (See Gibbons, 1988 for a more in-depth discussion.) Oosterbeek et al. implement these policies in the lab and find that workers' investment decisions are, at least on average, in line with theoretical predictions.[62]

61 For a review of the hold-up literature in the context of labor markets, see Malcomson (1997).

62 Sloof et al. (2007) consider the effects of keeping the level of one's specific investment secret on the hold up problem: the simple theoretical intuition is that if the other party does not know how much one has invested, they will be at an informational disadvantage in bargaining over the *ex post* surplus. The framing (and likely relevance) seems more appropriate to non-labor market interactions, however.

In sum, as for multitask principal-agent problems, experiments on multi-period principal-agent problems remain few in number. Nevertheless a few interesting results can be identified. They are:

(a) The early pooling equilibria at low effort levels predicted by ratchet effects models can be generated in the lab (Cooper et al., 1999).

(b) Consistent with Perfect Bayesian equilibrium in the modified game, labor market competition essentially eliminates the ratchet effect (Charness et al., 2008), at least in the case where workers' private information is about their ability (as opposed to the firm's technology). To our knowledge, no lab experiments on the latter ("hidden technology") variant of the ratchet effects model exist.

(c) The early signal-jamming equilibria at high effort levels predicted by career-concerns models can be generated in the lab (Irlenbusch and Sliwka, 2006; Koch et al., 2009).

(d) If agents choose efforts in both periods in a career-concerns game, the effects of making effort publicly observable sometimes contradict the career concerns model: Rather than reducing first-period effort (because signal jamming is no longer possible), making effort public actually raises first-period effort (Irlenbusch and Sliwka, 2006). This suggests that agents are attempting to signal some personal characteristic that is outside the model, such as "honesty" or a personal willingness to work hard.

(e) Even in situations where short-term contracts should theoretically guarantee workers the same return to firm-specific investments as long-term contracts, enforceable long-term contracts induce more worker investments in firm-specific skills (Anderhub et al., 2003).

(f) The nature of *ex post* wage bargaining (threat point versus outside offer), and promotion policies (up-or-stay versus up-or-out) affect investments in specific training (Oosterbeek et al., 2007a,b).

4. TOWARDS BEHAVIORAL PRINCIPAL-AGENT THEORY: FAIRNESS, SOCIAL PREFERENCES AND EFFORT

Perhaps the main contribution of experiments to principal-agent theory is the cascade of papers demonstrating the presence of "social preferences" (where one takes into account the payoffs, actions, and/or beliefs of other parties) in the laboratory. The essential content of social preferences is that people will deliberately sacrifice money to help other people or hurt other people, or even to keep their promises and thereby avoid guilt or a "cost of lying". The fact that people do not simply maximize their earnings has far-reaching consequences for theory; some efforts have been made in this regard (for example, see Von Siemens, 2004 for a characterization of optimal contracts with social preferences and hidden action).

The earliest experimental paper to convincingly demonstrate the existence of social preferences is the article on the ultimatum game in Güth et al. (1982). In this bargaining

experiment (which has been replicated hundreds if not thousands of times), one person is provisionally allocated a sum of money and chooses how much to offer a paired participant. If the offer is accepted, it is implemented; if it is rejected, both parties receive nothing. The main result is that people reject lopsided offers, even though it is costly to do so.

This result and many others like it have led to a number of models of utility. These can be roughly classified as either distributional or reciprocal, with some hybrid models (e.g., Charness and Rabin, 2002; Falk and Fischbacher, 2006). Distributional models presume that one cares about the payoffs of others regardless of the choices and perceived intentions of these others. There are two primary types of distributional models: "Difference-aversion models" (e.g., Loewenstein et al., 1989; Bolton, 1991; Fehr and Schmidt, 1999, and Bolton and Ockenfels, 2000) assume that players are motivated to reduce differences between their payoffs and others', while "social-welfare models" (e.g., Charness and Rabin (2002)) assume that people like to increase social surplus, caring especially about helping those (themselves or others) with low payoffs. On the other hand, reciprocity models (e.g. Rabin (1993), Dufwenberg and Kirchsteiger (2004)) assume that the desire to raise or lower others' payoffs depends on how fairly those others are behaving; in other words, how one perceives the intentions of other parties affects one's behavior.

In the field environment, it is difficult to rule out that behavior is driven by expectations of future material benefit, since there is typically repeated interaction in the field. In the laboratory, one can isolate social preferences by ruling out the possibility of future interaction, either by using a one-shot game or by anonymous re-matching. Of course, one must be concerned that the behavior found in the laboratory is specific to the laboratory. Regarding distributional preferences, charitable giving in the US (a notoriously individualistic society) exceeds 2% of GDP, with 90% of people donating; thus, most people are willing to contribute materially to the well-being of even anonymous strangers in the field. Regarding negative reciprocity, workers have been known to engage in sabotage or increased theft rates after a pay cut or other actions perceived to be unfair (see for example Greenberg (1990) and Schminki et al. (2002)), particularly when procedural justice in the organization is low (Skarlicki and Folger, 1997). And the studies by Krueger and Mas (2004) and Mas (2006, 2008) present results that show retribution with real firms and workers. The case for non-instrumental positive reciprocity in the field is weaker, but includes cases such as tipping when on the road or higher response rates to mailed surveys that include small gifts.

Nevertheless, institutional and contextual factors determine the extent that social preferences come into play. For example, it is conventional wisdom that inexorable forces drive out social preference in a market setting. Roth et al. (1991) find experimentally that competition essentially completely drives out fairness considerations when 10 people can potentially accept a demand (only one of the people who accept this demand is

selected) for a share of the pie that is made by one paired person. Recent papers such as Gneezy and List (2006) and Fershtman et al. (2009) suggest possible limitations for the applicability of social preferences. On the other hand, even though a great deal of research (see Holt, 1995 for a survey) has found a striking degree of convergence to the self-interested competitive equilibrium in experimental double auctions, Fehr and Falk (1999) demonstrate the presence of downward wage rigidity in an experimental labor market featuring the competitive double auction.

Fehr and Gächter (2000) suggest that a key determinant of whether social preferences come into play is whether contracts are complete and enforceable. They point out that incomplete contracts typically prevail in the labor market, where (for example) wages are often paid without any explicit performance incentives. They state: "In situations where contracts are reasonably complete, the underlying assumption of self-interest should continue to be especially important" (p. 178). However, based on evidence from Fehr et al. (2007b), where an implicit and incomplete contract with an unenforceable bonus outperforms a complete contract with contracted effort enforced through partial monitoring and the incomplete contract is selected 88% of the time, they argue that reciprocal considerations are not only critical with incomplete contracts, but also: "The endogenous formation of incomplete contracts through reciprocal choices shows that reciprocity may not only cause substantial changes in the functioning of given economic institutions but that it also may have a powerful impact on the selection and formation of institutions" (p. 178).

Given the evidence presented above, it seems clear that there is scope for improving the principal-agent models through the incorporation of behavioral motivations such as social preferences. Perhaps the insights into the nature of non-self-interested behavior gleaned from experiments will eventually be applied to a variety of economic settings, including employee response to changes in wages and employment practices.

4.1. Models of social preferences

We begin the discussion of the relevant literature by summarizing some of the more prominent models of social preferences. We proceed historically, rather than by publication date. For detail of the models and their full functional forms, we refer the reader to the individual papers.

Bolton (1991) develops a model in which people care about both their own money and their relative position; this model is based on the Ochs and Roth (1989) finding that people frequently make disadvantageous counter-offers in a two-round ultimatum game.[63] One receives negative utility from receiving less than the other person, but is not

[63] In their experiment, if a proposal is rejected there is another round in which the pie is substantially reduced; the person who rejected the initial offer can then make a proposal to split the smaller pie. A disadvantageous counter-offer is a proposal that, if accepted, would lead to a smaller material payoff for the second-round proposer than he or she had rejected in the initial proposal.

bothered by receiving more. There is a trade-off between one's material payoff and one's relative standing, in that a person might prefer (Own, Other) material payoffs of (3, 2) to material payoffs of (4, 6). This model is entirely consequential, as intentions or previous history does not affect preferences.

Rabin (1993) was the first to provide a formal game-theoretic model incorporating reciprocity preferences. The central notion is that of *kindness*, defined in terms of the payoff options made available to the other player by one's own choice, with one's kindness increasing as the best available payoff for the other person increases. If you are unkind, the other person may sacrifice money to hurt you, while if you are kind, the other person may sacrifice money to help you. In the latter case, mutual cooperation in the prisoner's dilemma can be supported as an equilibrium. Dufwenberg and Kirchsteiger (2004) formally extend this model (still including kindness as the main concept) to sequential games, which facilitates application to the more standard experimental games with first movers and responders. These models concentrate on reciprocity and only employ simplistic notions of fairness and distributional preferences.[64]

The Bolton and Ockenfels (2000) and Fehr and Schmidt (1999) consequential models of relative position (or inequity or difference aversion) were developed approximately simultaneously. The heart of these models is that people may (depending on parameter values) trade off money to make material payoffs closer together. The Bolton and Ockenfels model is more general and does not provide a specific functional form, while the Fehr and Schmidt model provides a simple and quite tractable functional form. The main difference between these models (in the two-player case) is that people are more bothered by being behind than by being ahead in the Fehr and Schmidt model, but that one's disutility from unequal payoffs is unaffected by whether one is ahead or behind in the Bolton-Ockenfels model.[65]

Falk and Fischbacher (2006) combine the Fehr and Schmidt notion of inequity aversion with reciprocity considerations in a complex hybrid model where a person is less bothered by another's refusal to come out on the short end of a split than by a refusal to share equally. Importantly, they assume that one does not resent harmful behavior by the other player if it seems to come only from the other player's unwillingness to come out behind rather than his or her selfishness when ahead. In other words, the prevailing social norm is one where it is bad form to be selfish when one has more than others, but is understandable when one has the short stick.

Charness and Rabin (2002) also combine distributional preferences with reciprocity.[66] The key innovation of this model is that people care about social efficiency, or the total payoffs for the reference group. Absent "misbehavior" (determined

64 However, there are two models in the appendix of Rabin (1993) that do consider the distribution of material payoffs.

65 In the n-player case, another difference is that Bolton and Ockenfels compare the ratio of one's material payoff to the total material payoffs, while Fehr and Schmidt consider the sum of pairwise comparisons.

66 Note that the real model is in the appendix of the paper.

endogenously), one's utility is determined by a weighted average of one's material payoff and a social component, which is itself comprised of a weighted average of the total payoffs for the reference group and the lowest payoff for anyone in the reference group. With misbehavior, there is negative reciprocity: one withdraws one's willingness to sacrifice to help the miscreant, by diminishing or eliminating the weight put on their payoff; one may even be willing to sacrifice to hurt the other person. However, in line with considerable experimental evidence (see the discussion below), there is no positive reciprocity in this model.[67]

Cox et al. (2007) present a non-equilibrium approach that combines a form of distributional preferences with reciprocity considerations. In this approach, both status (relative position) and reciprocity affect one's emotional state, which in turn affects the choices that are made by a utility-maximizing agent. They introduce a parametric model of other-regarding preferences in which one's emotional state determines the marginal rate of substitution between own and others' payoffs, and thus my subsequent choices. In turn, one's emotional state responds to relative status and to the kindness or unkindness of others' choices. Structural estimates of this model with six existing data sets demonstrate that other-regarding preferences depend on status, reciprocity, and perceived property rights.

A final model involves a trade-off between one's material payoff and one's degree of guilt from violating the expectations of another person. This is called "disappointment aversion" in Dufwenberg and Gneezy (2000), but is probably better known as "guilt aversion" (Charness and Dufwenberg, 2006). The idea is that the more one believes that a party with whom one is paired is expecting a favorable move, the more likely it is that one chooses the actual move. Charness and Dufwenberg (2006) find support for this notion, as there is a strong positive correlation between a responder's beliefs about the first-mover's beliefs and the responder's choice of the favorable action. Formal presentations of this concept ("simple guilt") and a more complex version ("guilt-from-blame") in which one feels guilt only to the extent that one believes that the other party blames one for an unfavorable outcome can be found in Battigalli and Dufwenberg (2007, 2009).[68]

In sum, a number of models of social preferences have been proposed, mainly classifiable into distributional models and reciprocity models and hybrids of these two approaches. Other motivations such as guilt aversion (and lying aversion, which has not generally been formalized; however, see Charness and Dufwenberg (forthcoming) for one approach) are also being considered in recent models.

[67] One might wonder whether a complex model such as Charness and Rabin (2002) is needed, as the substantially more parsimonious and tractable Fehr and Schmidt (1999) model does a reasonably good job of organizing much of the observed experimental behavior. While this topic is still being debated, it seems fair to say that there are many papers that provide data that cannot be explained by distributional models without considering reciprocity. These include Kahneman et al. (1986), Blount (1995), Charness (2004), Offerman (2002), Brandts and Charness (2003), Andreoni et al. (2002), Falk et al. (2003a,b), Charness and Rabin (2002), and Charness and Levine (2007).

[68] For another survey of social preferences, see Cooper and Kagel (in press).

(a) Distributional models posit that people are concerned with the payoffs of others in some fashion, with considerations of intention on the part of the others being irrelevant; in other words, only the consequences of one's choice matter. Bolton (1991) presumes that one cares about receiving less than another person in one's reference group, but doesn't mind receiving more than the other person. Bolton and Ockenfels (2000) instead presume that people care equally about coming out behind or coming out ahead of the average amount received by those people in the reference group. Fehr and Schmidt (1999) put forward a very tractable model that also presumes that people care about coming out behind or coming out ahead of others; however, one cares at least as much about coming out behind. In addition, one makes pairwise comparisons with the material payoffs of others, rather than comparing with the average of others (this only matters in environments with more than two people).

(b) Reciprocity models instead posit that people respond to the intentions of others, with perceived kind intentions being met with kind responses and perceived unkind intentions being met with unkind responses. In this way, both mutual cooperation and mutual defection are potentially equilibria in a Prisoner's Dilemma. These models typically abstract away from distributional considerations. Rabin (1993) is the seminal paper in this area, incorporating reciprocity into simultaneous games. Dufwenberg and Kirchsteiger (2004) extend this approach to sequential games, typically better-suited to the laboratory.

(c) Since both distributional and reciprocity considerations appear to be relevant, some recent models have combined these factors. Falk and Fischbacher (2006) combine intentions with Fehr–Schmidt preferences. Charness and Rabin (2002) put forth the notion of social efficiency, whereby people are typically interested in improving the payoffs of others, unless said others have behaved badly. There is negative reciprocity in this model, but (based on the experimental evidence in that paper and elsewhere) there is no positive reciprocity. Cox et al. (2007) use a non-equilibrium approach that does not require knowledge of beliefs and is therefore considerably more tractable.

(d) Other motivations such as guilt aversion have also been modeled (Battigalli and Dufwenberg, 2007, 2009). With guilt aversion, one trades off feelings of guilt against being selfish. The more one expects that another person expects her to behave favorably, the more likely one is to then behave favorably. Charness and Dufwenberg (2006) provide experimental support for guilt aversion, also finding that promises (statements of intent) are particularly useful in achieving optimal social outcomes.

4.2. The gift-exchange game

Probably no experimental game in the area of labor economics has had as much impact as the gift-exchange game, which tests the notion (Akerlof, 1982; Akerlof and Yellen, 1988, 1990) that there is a positive relationship between wages and effort. This game was designed to mimic an employment relationship, in that labor contracts are typically

incomplete and effort is not (fully) enforceable; thus a wage offer is binding, but effort is discretionary. In its basic form, the experimental participants are divided into "firms" and "workers", who interact anonymously either in some form of labor market or in one-to-one pairings; typically, the game is played for a number of periods. In this subsection we report results from the earliest gift-exchange experiments.

The first paper reporting results from the gift-exchange game is Fehr et al. (1993). They create a competitive labor market using a two-stage game. Workers and firms[69] (with more workers than firms) were separated into two rooms, and communication between the two rooms took place via telephone. The first stage was a one-sided oral auction with employers as bidders. Firms made wage proposals, which were posted in the room containing the workers. Once a worker accepted an offered wage, the first stage was concluded for both the firm and the worker. A firm could revise its (non-accepted) wage offer upward, so that it was higher than any existing posted wage offer. People who did not contract received zero earnings for the period. In the second stage, workers chose effort and this choice was only revealed to the paired firm. Firms chose wages (restricted to be multiples of five) from the interval [26, 126], earning (126-wage) × effort. Workers earned the wage less 26 less the cost of effort, shown below:

e	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$c(e)$	0	1	2	4	6	8	10	12	15	18

If workers are purely self-interested, the prediction is of course that they will choose $e = 0.1$ regardless of the wage. Knowing this, firms will choose a wage of 30, the lowest wage ensuring participation. However, the main result is that both wages and effort levels far exceed the predictions with self-interested workers, as the average wage was 72 and the average effort chosen was 0.4. In addition, there is a very strong positive relationship between effort and wage.

The second paper in this series is Fehr et al. (1998), who conduct both a one-sided oral auction treatment and a bilateral-gift-exchange treatment using Austrian soldiers to test the robustness of the Fehr et al. (1993) results and to determine the relative effect of competition on wages. An additional treatment involves complete contracts, in the sense that the experimenter enforced an effort level of 1.0 and no effort costs were subtracted from worker's earnings. The main finding is that gift exchange persists even in the absence of competition, when firms and workers are matched on a one-to-one basis; after a few periods wages in this treatment (and the ratio of effort to wage) coincide with wages in the one-sided oral auction without effort enforcement. Thus, the qualitative findings of Fehr et al. (1993) are replicated with non-student participants and even without competition. Wages were considerably lower when effort is enforced,

[69] The language used in this paper involves buyers and sellers, prices and quality; however, the paper emphasizes the labor interpretation, so we adopt the terminology used in almost all subsequent papers.

although still substantially above the minimum needed to ensure participation (workers could reject offers).

Fehr et al. (1997) study the impact of reciprocity on contract terms and their enforcement with three experimental treatments involving competitive markets and more workers than firms. In each treatment, the firm specifies a wage, a desired effort level, and a fine imposed if the firm detects that the worker has shirked (provided less than the contracted level of effort). In the weak-reciprocity treatment, workers who have accepted posted contracts choose effort levels, and a random device determines whether shirking is verifiable (and fined at the specified level); the firm then learns the chosen effort level. The no-reciprocity treatment is identical, except that the experimenter exogenously fixes the effort level. Finally, in the strong-reciprocity treatment, there is an additional stage in which firms can reward or punish workers at a cost. The results show that firms demand and enforce much higher effort levels in the strong-reciprocity treatment than in the weak-reciprocity treatment; there is much less shirking in the strong-reciprocity treatment. In fact, both firms and workers earn more in the strong-reciprocity treatment, in large part because the higher effort levels lead to a larger pie. Nevertheless, firms' contract offers are much higher in the weak-reciprocity treatment than in the no-reciprocity treatment, and the offers increase with the desired effort level.

In sum, the gift-exchange game has been a very successful approach to modeling labor issues in the laboratory. The main finding is that higher wages lead to higher effort. This section reports only the earliest gift-exchange experiments.

(a) Some early gift-exchange papers used a form of one-sided oral auction to create a market environment. Fehr et al. (1993) have more workers than firms, simulating unemployment conditions. Fehr et al. (1998) find very similar results with Austrian soldiers and a bi-lateral design, where there are equal numbers of workers and firms.

(b) Fehr et al. (1997) find that both firms and workers earn more when firms can punish or reward workers for their effort choices, since higher effort is socially efficient (there is a larger pie to divide). Even though contract offers are higher without this enforcement possibility, there is much more shirking (low effort). Nevertheless, offers increase with the desired effort level in both cases.

4.3. Multi-worker gift-exchange experiments

It seems substantially more realistic to consider an environment in which a firm has more than one employee. When there are multiple workers who receive wages and who can provide effort that benefits the firm, there is the possibility of dispersed responsibility for the firm's earnings (leading to possible free riding) and there are considerations of horizontal fairness (one's pay compared to the pay of other workers).

Maximiano et al. (2007) compare a bilateral gift-exchange game with one in which each firm has four workers; in the latter case the employer is likely to earn much more than any of her workers, thus reducing the need for any individual worker to sacrifice to

help the relatively high-income firm. The authors did not expect gift exchange to survive in this environment.[70] In fact, effort levels in the latter treatment are only marginally lower than in the bilateral game, so that "the gift exchange relationship is quite robust to increases in the size of the workforce" (p. 1026).[71] Their results suggest that intentions-based reciprocity is a driving factor, although efficiency preferences (Charness and Rabin, 2002) may also play a role in inducing this behavior.

Charness and Kuhn (2007) match two workers (with different productivity levels, although the precise levels are unknown to the workers) with one firm to investigate whether workers have concerns with pay inequality and whether pay secrecy and pay compression is therefore beneficial for a firm; in a within-subjects design, we varied whether a worker was aware of the other worker's wage. Under fairly general conditions, we demonstrated theoretically that workers' responsiveness to co-workers' wages should lead profit-maximizing firms to compress wages or maintain pay secrecy. And, as in other gift exchange experiments, we observed a strong positive empirical relationship between "own" wages and effort. Surprisingly, however, the effort level provided by a worker was unaffected by the wage paid to his or her co-worker. Furthermore, although firms compress wages when the co-workers will know both wages, this did not raise profits. It seems that the relationship between a worker and the firm is much more salient than the relationship between the pay of the co-workers. Overall, our experimental evidence "casts doubt on the notion that workers' concerns with equity might explain pay policies such as wage compression or wage secrecy" (p. 693).

Gächter et al. (2008) perform an experimental analysis of pay-comparison information and effort-comparison information in an environment in which firms are matched with two workers. While effort is highly sensitive to the worker's wage, co-worker wages *per se* have no effect. Further, when the firm pays different wages to the workers, the co-worker's effort decision is ignored. However, worker behavior is affected when both pieces of social information are provided: a generous wage generates higher effort when one's co-worker exerts high effort, but is ineffective when the co-worker contributes little or no effort. They suggest that group composition is a relevant factor for obtaining beneficial effects from social information.

Abeler et al. (forthcoming) focus on a two-worker-one-firm environment in which a worker knows both the ability and effort level of the co-worker. However, they reverse the order of play, as workers first choose effort levels and the paired firm then chooses

[70] They state on p. 1026: "Frankly, we did not believe that the gift exchange model would survive the more realistic multiple workers per employer design. We were wrong".

[71] We hasten to add that, despite the large number of papers that find a strong positive relationship between wage and effort, gift exchange is not robust to all experimental conditions. Hannan et al. (2002) find little gift exchange with undergraduates; on the other hand, MBA students provide effort responses similar to those in most gift-exchange studies. Charness et al. (2004) find that when a complete payoff table is provided in the experimental instructions, workers choose substantially less effort compared to a treatment in which participants had only the information needed to compute their payoffs.

wages. In one treatment firms are constrained to pay equal wages, while in a second treatment there is no such constraint. Perhaps surprisingly, they find that there is lower effort when the firm is forced to pay equal wages, as a worker who chooses higher effort than the co-worker does not feel it is fair to receive the same wage as the co-worker, and subsequently reduces effort. Thus, it is important to pay attention to equity considerations rather than equality *per se*.[72]

In sum, there have been a modest number of gift-exchange experiments with multiple workers, a more realistic case than the standard game.

(a) Maximiano et al. (2007) find that gift exchange is relatively undiminished even when each firm has four workers, with their results suggesting that intentions-based reciprocity is a driving factor in the effort choices.

(b) Some recent papers have examined the effect of pay-comparison information and/or effort-comparison information when there are two workers per firm. Both Charness and Kuhn (2007) and Gächter et al. (2008) find that pay-comparison information alone has little or no effect on the effort choices of workers. However, the latter study finds that worker behavior is affected when both pay-comparison information and/or effort-comparison information are provided and the co-worker provides high effort.

(c) Abeler et al. (forthcoming) examine the two-worker-per-firm environment in which a worker knows both the ability and effort level of the co-worker, with workers first choosing effort and the firm then choosing the wage. Effort is lower when firms are forced to pay the same wage, due to workers who choose higher effort objecting to being paid the same as a shirking co-worker.

4.4. Positive and negative reciprocity

While the classic gift-exchange experiments provide strong evidence of reciprocal behavior, it is less clear that this represents reciprocity in a strict sense. The reason for this is that people may have distributional preferences that could lead to the same behavior that is observed in an environment where reciprocity is possible. Thus, positive or negative reciprocity reflects behavior that differs from what a responder would have done in the absence of a first-mover action that is perceived to be positive or negative. Some experimental games from Charness and Rabin (2002) illustrate this point. In the baseline case, a participant unilaterally chooses between (Other, Own) material payoffs of (400, 400) or (750, 375). Around fifty percent of the population chooses to sacrifice 25 units to give the other person an additional 350 units, even though this leads to a large difference in material payoffs.

In a second case, the other paired participant first faced a choice between payoffs of (750, 0) or passing the choice to the second participant, who would once again face a

[72] In a related paper, Meidinger et al. (2003) examine team heterogeneity and productivity. There is considerable free riding as each agent is greatly influenced by his or her teammate's behavior. On the other hand, workers are better able to cooperate when the team is homogeneous.

choice between (Other, Own) payoffs of (750, 375) or (400, 400). Positive reciprocity would imply that the rate of (750, 375) choices should increase, as the first mover would clearly seem to be kind by allowing the responder to receive a positive payoff. However, in fact the rate goes down slightly (to 39%)! So positive reciprocity doesn't seem to be present here. In a third case, the other paired participant first faced a choice between payoffs of (550, 550) or passing the choice to the second participant, who would once again face a choice between (Other, Own) payoffs of (750, 375) or (400, 400). Negative reciprocity would imply that the rate of (750, 375) choices should decrease, as the first mover would clearly seem to be unkind by forcing the responder to receive a smaller material payoff than was available with the outside option. In fact, the rate of (750, 375) choices does decrease sharply to 11%; thus, negative reciprocity is a factor.[73]

Perhaps the first experimental paper to carefully test for positive and negative reciprocity was Charness (2004).[74] This paper considers gift-exchange in a bilateral setting, varying whether the wage was determined by a self-interested firm or generated by an exogenous process (such as a draw from a bingo cage); in all cases, the firm benefits from the worker's chosen effort. There is a strong positive relationship between wage and effort in all treatments. However, the effort level with low wages is lower when the wage was chosen by a self-interested firm than when it was generated exogenously, suggesting the presence of negative reciprocity. On the other hand, there was virtually no difference across treatments in the effort level with high wages. Thus, this paper was the first to provide experimental evidence that positive reciprocity seems to be much weaker than negative reciprocity, while at the same time clearly identifying the effect of the distribution of payoffs *per se* on behavior.

A subsequent and related paper (in a non-labor setting) is Offerman (2002), who studies the effects of random choice mechanisms while allowing for both positive and negative reciprocity. He considers players' responses to a helpful or hurtful choice, as a function of whether the "choice" was made by an interested party or generated at random (responders could sacrifice one unit to either increase or decrease the first-mover's payoff by four). Following the helpful choice, responders never paid to lower the first-mover's payoff, but paid to help first movers more often when the first mover made the choice than when the choice was randomly-determined. This suggests some positive reciprocity may be present, although the effect is not significant. On the other hand, the effect on the response to the first-mover's perceived intentions was dramatic and significant following a hurtful choice, again suggesting that hurting hurts more than helping helps (the title of the paper).

73 A further example from Charness and Rabin (2002) provides evidence of negative reciprocity. In ultimatum-game experiments, a high proportion of responders reject offers that would give them only 20% of the pie. However, when we offered 36 participants a choice between (Other, Own) payoffs of (800, 200) or (0, 0), exactly zero chose (0, 0). Thus, distributional considerations are insufficient to explain rejections in the ultimatum game, so that negative reciprocity would appear to be the driving force.

74 Once again, we order papers by the date they were written (in this case, 1996), rather than the publication date.

Brandts and Charness (2003) test for punishment and reward in a cheap-talk game and show that intention is a critical issue, finding substantial negative reciprocity and limited positive reciprocity. One player sends a message about her intended play to another player; after play takes place, the other player is then given an opportunity to punish or reward the first player. The authors found that the responder was twice as likely to punish unfavorable play by the first player if that first player had lied about his play than if he had told the truth. A relatively small number of responders chose to reward a favorable play by the first mover.

On the other hand, Cox (2004) reports significant positive reciprocity in the investment game (Berg et al., 1995).[75] The triadic design compares behavior in the standard game with behavior when the first mover is a dictator and to behavior when the experimenter determines the amount received by the responder rather than this being determined by a self-interested first mover; this procedure should allow one to distinguish distributional preferences *per se* from reciprocal preferences. Cox et al. (2008) finds significant positive reciprocity, but does not find significant negative reciprocity in the "moonlighting game", where the first mover can take from the responder as well as pass to the responder. Finally, Cox and Deck (2005, 2006) report mixed findings on positive reciprocity; the results depend on whether or not the experimenter can observe the actions of the players (double blind or single blind), as well as on the sex of the subject.

In sum, the experimental evidence regarding intention-based positive and negative reciprocity is mixed, although the general result is that negative reciprocity is stronger than positive reciprocity. In a sense, this can be seen as reflecting expectations and violations thereof, as seen in a self-serving way. If one expects kind or favorable treatment and receives it, there is no strong emotional jolt; on the other hand, if one has this expectation and receives unkind or hurtful treatment, the emotional response is much stronger.

(a) Negative reciprocity is highly pervasive in experiments, much as it is in the field.
(b) On the other hand, a number of papers, such as Charness (2004), Offerman (2002) and Charness and Rabin (2002) find little or no evidence of intentional positive reciprocity, in the sense that responses after a kind action do not receive a more favorable response than when no action has been taken.
(c) However, this is not a universal result, as papers such as Cox (2004) and Cox et al. (2008) report significant positive reciprocity. Still, Cox and Deck (2005, 2006) report mixed findings on positive reciprocity. So this topic remains open.

4.5. Pay regulation

Governments often consider whether to regulate pay by means such as mandating a minimum wage or a sick-pay rate. What are the effects of such policies? In the field,

[75] In the standard version of this game, both a first mover and a responder are endowed with 10 units. The first mover can pass up to 10 units, with the amount passed tripled by the experimenter and then received by the responder. The responder can then pass back any number of units (not tripled).

it is difficult to ascertain this, as there are many factors present and changing at the same time. Thus, experimental techniques are likely to be useful, as the effects of such policies can be successfully isolated.

Brandts and Charness (2004) were the first to study the effect of a minimum wage in a gift-exchange format. They created a more symmetric payoff design. Each person was endowed with 10 units, the first mover could pass up to 10 units, and the responder received five times whatever was passed; the responder could then send back up to 10 units, with the original first mover receiving five times whatever was passed back. Brandts and Charness imposed a minimum wage of five in a condition with an excess supply of workers. The mandated minimum wage was counterproductive in the sense that the average effort was 30% lower than without it, even though the average wage was 5% higher. In addition, they found that the highest wage was chosen only half as frequently in the minimum-wage condition, with effort also reduced by 30% at this top wage.

Falk et al. (2006) test the effect of imposing a minimum wage in an environment in which each firm is matched with three workers and must choose the same wage for each of the workers. The workers decide on the minimum wage that they would conditionally accept (the strategy method) and the actual assigned wage is then accepted or rejected accordingly; however, workers do not choose effort levels, and a firm's profits only depend on the number of workers who accept the wage offer. In some sessions, fifteen periods of a mandated minimum wage were followed by 15 periods with no minimum wage; the order was reversed in the other sessions. They find that there are lasting consequences of the mandated minimum wage even after it has been removed, in that firms must pay higher wages after the removal than before it was imposed. Thus, the authors conclude that policy can affect people's sense of what is a fair wage.

Owens and Kagel (2010) use a variant of the payoff design in Brandts and Charness (2004). They find that an imposed minimum wage reduces effort in the neighborhood of the minimum wage, but that there are no significant effects on effort levels for higher wages. The minimum wages leads to improved incomes for both firms and workers (particularly the latter). Since there is little effect at higher wages, it appears that the minimum wage requirement is less salient at these higher wages and/or that employees recognized that wages set a good deal higher than the minimum represent just as large a monetary gift as without the presence of the minimum.

Dürsch et al. (2008) report the results of a form of gift-exchange game that investigates the issue of sick pay in an experimental labor market. The main variation is whether there is one-to-one matching of firms and workers or whether there is a market setting in which firms post wage offers and workers select the ranking of their preferences regarding the possible offers; in the latter case, a firm may end up hiring more or less than one worker. In all cases, there is a one-third chance (chosen at random) that a worker will be "sick" and so be unable to complete his or her intended effort. Firms chose one of

five possible contracts, each of which specified a wage to be paid in case of zero effort (not showing up for work, either because the computer made this choice or because the worker chose zero effort voluntarily; the firm cannot tell if the worker is actually sick or simply chose to stay home) and a wage to be paid in case of positive effort. The main findings are that higher wage offers significantly increase effort choices (both with respect to sick pay and non-sick pay), firms can attract more reciprocal workers by offering sick pay, but firms benefit from offering sick pay only in the market setting where there is competition for workers.

Bauernschuster et al. (2009) use a similar design to consider the effect of sick pay with heterogeneous workers with different likelihoods of being "sick". They use a 2×2 experimental design, in which they vary whether there is a minimum sick-pay rate (40% of the wage) and whether workers all have the same 20% rate of being sick or whether half of the workers have a 10% chance of being sick and the other workers have a 30% chance of being sick. Issues of interest include the degree of moral hazard (pretending to be sick), whether the adverse-selection problem is severe enough to lead to a collapse in the market for sick pay, and whether higher levels of sick pay lead workers to choose higher effort levels. The main results are that higher-risk workers do indeed select into contracts with higher sick pay rates, and they pay for this by getting lower wages and an overall worse deal than the low risk workers, higher sick pay leads to increased effort, but only in the case where the sick pay rate is freely-chosen, and the sick-pay market does not break down due to the adverse-selection problem.

In sum, there are mixed results regarding the effect of imposing a minimum wage. Regarding sick pay, it appears it is possible (but perhaps not easy) to design systems where this is a useful policy intervention.

(a) Brandts and Charness (2004) find that a mandated minimum wage was counter-productive in terms of effort provision. High wages were not offered nearly as frequently when this mandated minimum wage is present, as it appears that intrinsic generosity or the perceived need for high wages is crowded out by the mandated minimum wage.

(b) On the other hand, Falk et al. (2006) find that a minimum wage is potentially beneficial. They find lasting positive consequences from the mandated minimum wage even after it has been removed, as firms higher wages after the removal than before it was imposed. Owens and Kagel (2010) also find some beneficial effects from imposing a minimum wage, with Pareto improvements resulting for firms and workers.

(c) Dürsch et al. (2008) and Bauernschuster et al. (2009) study the effects of sick pay provision in an experimental labor market. In the first paper, higher wage offers significantly increase effort choices (both with respect to sick pay and non-sick pay), but firms only benefit from offering sick pay when there is competition for workers. In the second paper, workers with a greater likelihood of being "sick" select into contracts with higher sick pay rates, but at a cost in terms of wages. Higher sick pay

can result in increased effort in some circumstances, and the adverse selection problem does not destroy the sick-pay market.

4.6. Do gift-exchange and social preferences map into the field?

A major question that labor economists might have concerning laboratory gift-exchange results is the degree to which these mean anything in the field environment (external validity). Two recent papers call into question the degree to which the social preferences exhibited in these experiments are robust. On the other hand, a number of other studies provide evidence that the social preferences identified in the laboratory map well into the real world.

Gneezy and List (2006) investigate whether gift-exchange behavior persists over time in two real-effort tasks, one involving work in a library and the other involving door-to-door solicitation. The mechanism for testing for reciprocity consists of the experimenters telling people that they will be receiving a certain piece rate for their work, but then announcing at the time of the six-hour task that in fact they will be paid a substantially higher piece rate. The results are persuasive, particularly for the fundraising task: people do in fact work harder for the surprise pay, but this effect vanishes over the course of time. Thus, these results constitute a cautionary note against applying the results of these laboratory experiments to the field environment. However, one caveat is that these experiments only pertain to *positive* reciprocity, as even the lower advertised pay rate is above the alternative wage that the student workers would normally earn. Given the relative dearth of positive reciprocity in laboratory experiments, it may not be so surprising that positive reciprocity in these field experiments is fleeting.

Fershtman et al. (2009) present compelling evidence that social preferences are subject to framing effects and are not general to all environments. They show that introducing a competitive frame to the environment crowds out social preferences in mini-dictator and trust games. For example, 72.5% of dictators choose an (Own, Other) allocation of (8, 8) instead of (11, 2). This is followed by a real-effort competition between the dictator and the recipient, where the (8, 8) outcome is implemented if the recipient wins but the (11, 2) outcome is implemented otherwise. If the dictator really wanted the (8, 8) split, the simple strategy would be to solve no problems; however, dictators contribute considerably more effort when solving more problems leads to higher own payoffs than when solving more problems doesn't affect these payoffs (in the baseline condition). Furthermore, in another treatment the (11, 2) split is implemented if the dictator loses to the recipient, while otherwise the (8, 8) split is implemented. In this case, 85% of the dictators do nothing. This result demonstrates that social preferences may not have much effect when competition is salient.

Two other experiments present a different view than Gneezy and List (2006). Bellemare and Shearer (2009) conduct a field experiment that investigates worker responses to a monetary gift from their tree-planting firm using incentive contracts.

Workers were told that they would receive a pay raise for one day. Productivity on the day of the gift is compared with productivity on adjacent days, under similar planting conditions. They find direct evidence of a significant and positive effect on daily planter productivity on the day of the gift, controlling for a variety of other possible factors. Kube et al. (2006b) conduct a field experiment involving a six-hour task in which the participants were told that they would presumably receive 15 Euro per hour; in three treatments, they were then paid either 10, 15, or 20 Euro per hour. The main result is that there is little difference in performance in the treatments where people are paid either 15 or 20 Euro per hour, but that there is a strong, deleterious, and lasting effect on performance when people are paid less than the presumptive rate. These results suggest that there is an asymmetry between positive and negative reciprocity in the field, very much in line with the evidence found in laboratory experiments. However, one caveat to these results is that the sample size is quite small.[76]

There are two main threads of evidence suggesting that social preferences are linked to the general population and the workforce. First, there are a number of non-laboratory experiments in European nations that show that the basic behavioral patterns observed in fairness-related experiments with students also prevail in representative samples of the general population, with older cohorts generally being more reciprocal than younger cohorts.[77] Holm and Nystedt (2005) analyze behavior in the investment game, with people selected from a public database in Sweden. While the average amount returned was similar for both cohorts, the proportions dispersed for the older responders, suggesting a greater degree of responsiveness to the environment. Other studies conducted with representative surveys in Germany (Fehr et al., 2002) and the Netherlands (Bellemare and Kröger, 2007) conclude that older cohorts are more generous as responders. Falk and Zehnder (2006) find substantial evidence of social preferences in a sequential trust game conducted with 1000 residents of Zurich, also finding evidence of discrimination and in-group favoritism. Sutter and Kocher (2007) find that the elderly are more reciprocal, in a study with participants ranging from 8-year-old children to people in their late sixties. In their study, trust increases from early childhood to early adulthood, but stays constant thereafter. Finally, Karlan (2005) finds that the amounts returned in a laboratory investment game conducted in Peru predict micro-finance loan payments a year after the experiment.

The second thread involves studies that show a relationship between worker productivity and social preferences. Barr and Serneels (2009) conducted a study with Ghanaian manufacturing workers. They find that a measure of trustworthiness (the ratio of the amount returned to the amount sent) has a positive relationship with the wages

[76] In a related study, Kube et al. (2006a) conduct a controlled field experiment that tests the extent to which cash and non-monetary gifts affect workers' productivity. The main result is that a non-monetary gift leads to a significant and substantial increase in workers' productivity, while a cash gift of the same value is ineffective.

[77] Charness and Villeval (2009) also find that students are less cooperative than young workers, older workers, and retirees.

paid, with these wages being used as a proxy for productivity. This is particularly true with respect to the average trustworthiness in the workplace in question, although the direction of any causal relationship is unclear—are people more trustworthy because their wages are higher or *vice versa*? They conclude that behavioral characteristics and corporate culture are major determinants of the productivity of the firm and the operation of the labor market. Carpenter and Seki (forthcoming) conducted a series of experiments with local shrimp fisherman in Japan. They used a laboratory public-goods game to measure social preferences and also obtained measures of fishing productivity. The main finding is that there is a strong correlation between measured social preferences and productivity for the fisherman; a second finding is that social preferences grow with the degree to which team production is present in the fishing environment.

In sum, there is mixed evidence regarding the extent to which social preferences manifest in the field. It is clear that these are not present in all environments, but there are a number of papers providing evidence from laboratory and field experiments, as well as work environments and surveys, that suggest that social preferences can be a factor in the field.

(a) Gneezy and List (2006) find positive reciprocity in the first three hours of field experiments, but not in the last three hours. They interpret their results as indicating that social preferences are ephemeral. Similarly, Fershtman et al. (2009) use a clever design to show that social preferences are crowded out when a competitive game is played after an initial choice. Clearly social preferences are not ubiquitous.

(b) However, other studies find rather different results. Bellemare and Shearer (2009) present direct field-experimental evidence of a significant and positive effect on productivity on the day when a gift is given, controlling for a variety of other possible factors (yet it remains unclear how long such an effect will persist). Kube et al. (2006b) use a design in a field experiment that can test for both positive and negative reciprocity. In line with many experimental results, there is little evidence of positive reciprocity, but strong evidence of negative reciprocity when the pay rate is lower than what was anticipated.

(c) Survey evidence indicates that social preferences are linked to the general population and the workforce. Holm and Nystedt (2005), Fehr et al. (2002), Bellemare and Kröger (2007), Falk and Zehnder (2006), Sutter and Kocher (2007), and Karlan (2005) provide such evidence from surveys in several European nations, as well as Peru. In general, it seems to be that older people are more reciprocal than younger people, suggesting that this is a learned trait.

(d) There is also direct evidence from the work environment. Barr and Serneels (2009) find a relationship between worker productivity and social preferences for Ghanaian manufacturing workers. Carpenter and Seki (forthcoming) observe a definite positive relationship between social preferences and productivity for shrimp fishermen in

Japan, with said social preferences increasing to the extent that team production is present in the fishing environment.

4.7. Communication

In this final subsection, we consider how communication can affect behavior in principal-agent relationships. While it is true that other forms of social considerations such as distributional preferences and negative reciprocity are likely to lessen the applicability of the equilibrium contracts derived from standard theory, some studies demonstrate that communication can lead to better social outcomes than can be achieved with more standard social preferences.

Charness and Dufwenberg (2006) examine experimentally the impact of communication on trust and cooperation with hidden action (moral hazard). The principal has an outside option that (in the main treatment) gives (Principal, Agent) payoffs of (5, 5); alternatively, the principal can leave matters up the agent, who chooses between (Principal, Agent) payoffs of (0, 14) and expected payoffs of (10, 10).[78] With standard preferences, the agent would choose (0, 14), so that the principal will choose (5, 5). In fact, there are substantial rates of cooperation even in the absence of communication, but the likelihood of the (10, 10) expected outcome increases from 20% to 50%. Furthermore, this increase is almost entirely driven by free-form and endogenous *promises* (statements of intent) by the agent to behave cooperatively. The evidence is consistent with people striving to live up to others' expectations so as to avoid guilt. When players exhibit such guilt aversion, communication may influence motivation and behavior by influencing beliefs about beliefs. Charness and Dufwenberg argue that guilt aversion may be relevant for understanding strategic interaction in a variety of settings, and that it may shed light on the role of language and social norms in these contexts.

Brandts and Cooper (2007) study manager-worker interactions in an environment where payoffs depend on employees coordinating at high effort levels; the lowest effort level chosen by any worker determines the overall production. One treatment investigates the effect of increased incentives (the marginal benefit of coordinating on a higher effort level increases substantially) on worker behavior, while a second treatment allows the manager to communicate via chat with the workers. It turns out that communication is a more effective tool than incentive changes for improving coordination on high effort levels. An analysis of the content of communication indicates that the most effective communication strategy is to request high effort, pointing out the mutual benefits of high effort. Thus, direct financial incentives have some benefit in terms of coordination, but messages are the better strategy if one cannot choose both.

Charness and Dufwenberg (forthcoming) investigate the impact of communication on trust and cooperation with hidden information (adverse selection). There are two

[78] In fact, there is a 5/6 chance that these payoffs are (12, 10) and a 1/6 chance that they are (0, 10). In this way, the principal cannot be certain of the agent's choice when a zero payoff is received.

possible agent ability levels, with a 1/3 chance that an agent will have high ability. The principal has an outside option or can leave matters up the agent. There are two main cases. In the first, there are possible Pareto-improvements over the outside option for both types of agents. In the second case, there is no feasible Pareto-improvement for the low-ability agent. While the standard game-theoretic prediction is for the principal to choose the outside option in both cases, the authors find a dramatically different effect of free-form communication from the agent to the principal in these two cases. When a Pareto-improvement is feasible with a low-ability agent, communication doubles the rate at which the low-ability agent sacrifices money to help the principal. However, communication has no effect in the alternative case. The difference is driven by the fact that many messages in the first treatment confess that the agent has low ability, but that he or she will choose the Pareto-improvement over the selfish option (and they all do so). Charness and Dufwenberg conclude that is good policy to offer lower-ability workers an opportunity to participate in a socially-beneficial outcome, as they are likely to behave cooperatively.

In sum, while experiments on cheap talk have been conducted for over 20 years, some recent work has applied such nonbinding pre-play communication in clear principal-agent environments. This is a very promising area for both future research and policy.

(a) Charness and Dufwenberg (2006) find that promises (statements of intent) by the agent to behave cooperatively are very effective in achieving optimal social outcomes in a hidden action (moral hazard) setting, where the standard prediction is a total lack of cooperation. The evidence supports the notion of guilt aversion, where people strive to avoid guilt stemming from disappointing the expectations of others.

(b) Brandts and Cooper (2007) find the striking result that coordination on higher effort is facilitated more by communication between a manager and the workers than by increased incentives. Communication about the mutual benefits of high effort is particularly valuable. Thus, while direct financial incentives may be somewhat useful, communication in a team-production environment is even more so.

(c) Charness and Dufwenberg (forthcoming) also find that promises are effective in a hidden information (adverse selection) environment, but only when less able agents can participate in a Pareto-improvement in material payoffs. A potential policy application is to offer lower-ability workers an opportunity to participate in a socially-beneficial outcome, as they are likely to behave cooperatively.

5. MORE LAB LABOR: BARGAINING, SEARCH, MARKETS, AND DISCRIMINATION

In this section we summarize the contributions of laboratory experiments to aspects of labor economics other than the principal-agent problem or labor supply issues. These are, in turn, bargaining, strikes and arbitration; search; models of labor market equilibrium; and the study of gender, race and discrimination in labor markets.

5.1. Bargaining, strikes and arbitration

The bilateral bargaining problem—how a surplus is divided between two players—has a long and rich history in economic theory, dating back at least to Nash's (1950, 1953) contributions to co-operative and non-cooperative bargaining, and important subsequent developments such as Rubinstein (1982).[79] Experimental studies of bargaining also have a long history, both in psychology (see for example Deutsch and Krauss, 1960) and economics (Siegel and Fouraker, 1960). Roth (1995) reviews the experimental economics literature on bargaining up to that point, much of which aims to test predictions of the Nash, Rubinstein and related models. As already noted, a well known and robust result of these tests is that bargaining outcomes in the lab are not well described by perfect equilibria of games between rational, selfish agents; behavior is often highly suggestive of social preferences among the bargainers.

Given these early laboratory findings, it is perhaps not surprising that much of the more recent experimental bargaining literature has attempted to understand the nature of apparent social preferences in bargaining situations; this is typically done using extremely simple bargaining games (such as the ultimatum and dictator games) where there are fewer confounding factors, including strategic considerations, that might also affect outcomes. Recent examples of this approach include Charness and Rabin (2002), Andreoni et al. (2003), and Falk et al. (2003a,b); for reviews see Hoffman et al. (2008) and Murninghan (2008b). Other aspects of the bargaining problem that have been examined using laboratory experiments since Roth's review include the effects of gender and other demographic differences (Eckel and Grossman, 1997; Holm and Engseld, 2005); of cultural differences (Chen and Tang, 2009); stakes (Slonim and Roth, 1998; Cameron, 1999; Munier and Costin, 2002); risk aversion and beliefs (Dickinson, 2009); deadlines (Sterbenz et al., 2001; Gneezy et al., 2003; Guth et al., 2005); incomplete information (Guth et al., 1996); face-to-face interaction (Ockenfels and Selten, 2000); delegation (Schotter et al., 2000); "hot" versus "cold" decision making (Brosig et al., 2003); communication (Brosig et al., 2004; Croson et al., 2003); self-serving biases (Babcock and Loewenstein, 1997); outside options (Carpenter and Rudisill, 2003) and situations involving bargaining among more than two players (Frechette et al., 2005; Charness et al., 2007).

While the above bargaining models and experiments were motivated and framed in a variety of ways (including relatively abstract frames, or as buyer-seller interactions), labor economists have found them of primary interest for the light they might shed on union-firm negotiations.[80] In particular, labor economists have used formal bargaining models and laboratory experiments to shed light (a) on the determination of disputes

[79] Summaries of these theoretical developments are available in Binmore and Dasgupta (1987), and in Osborne and Rubinstein (1990), as well as many other sources.

[80] Another important and promising application of bargaining experiments of interest to labor economists is to gender differences in labor market outcomes (e.g. Eckel and Grossman, 1997; Babcock and Laschever, 2003).

(strikes, lockouts, holdouts) and (b) on mechanisms—in particular, arbitration—that are designed to reduce the costs of such disputes.[81] We consider these two applications of bargaining theory and experiments in turn below.

While economists have been discussing the determinants of strikes since Hicks (1932) stated his famous "paradox" (if parties can forecast the bargaining outcome after a strike they can always do better to settle on that outcome immediately without a strike), theoretical and experimental analysis of strikes by economists did not begin in earnest until the advent of asymmetric information models of bargaining, which produced disputes such as strikes as equilibrium outcomes. Early theoretical contributions of this nature include Hayes (1984); Kennan (1987) provided an early review of the "economics of strikes". Since then, Sopher's (1990) laboratory experiment tested the simple "joint cost" theory that strikes will be less common when their social cost to the bargainers is higher in a shrinking pie bargaining game. Forsythe et al. (1991a,b) studied "strikes" in similar games with one-sided private information. Consistent with other laboratory bargaining experiments, they found that agents often forgo rewards in order to inflict punishments, and engage in other apparently fairness-driven behavior. In (1991b) the same authors describe a symmetric-information experiment. Kennan and Wilson (1993) review the literature on bargaining under asymmetric information up to that point, including the contributions of lab experiments. Since then, a few papers have used simple models of bargaining and laboratory experiments to address an old question in industrial relations: the effects of centralized versus decentralized bargaining on strikes and other outcomes (Berninghaus et al., 2001, 2003; Tournadre and Villeval, 2004).[82]

The aspect of bargaining on which labor economists and others have conducted the largest number of laboratory experiments is not strikes *per se*, but the effectiveness of dispute resolution (i.e. arbitration) mechanisms in encouraging bargaining and reducing disputes. In at least one way this is not surprising, as arbitration is a classic example of mechanism design, and a laboratory experiment is a natural way to "pre-test" a mechanism under ideal conditions to see if it has a reasonable chance of functioning as predicted in the field.[83] In addition, there is of course considerable policy interest in the design of arbitration mechanisms that encourage (efficient) settlements and minimize the use of the mechanism itself (especially in view of Crawford's 1979 dictum that the best arbitration system is one that is never used). Thus, the evaluation of arbitration mechanisms is one area where laboratory research by labor economists has already been successful in influencing economic policy.

81 There is also a literature that models the wage and employment outcomes of union-firm bargaining, and attempts to test for the appropriate model of the bargaining process. For reviews, see Farber (1987) and Kuhn (1998); we are not aware of any laboratory experiments in this area.

82 Non-experimental studies of strikes motivated by asymmetric information bargaining models include Card (1990), Cramton and Tracy (1992), Gu and Kuhn (1998) and Kuhn and Gu (1999).

83 Labor economists have also been active in the use of field data to study the effects of arbitration; see for example Currie and McConnell (1991), who studied dispute rates for public employees across Canadian provinces with different arbitration laws.

While laboratory experiments testing different arbitration (and other forms of third-party intermediation) systems have a long history in psychology (see for example Johnson and Tullar, 1972), the earliest experiment on arbitration by economists of which we are aware (Farber and Bazerman, 1986) is noteworthy for its methodology in view of current debates in economics about the lab "versus" the field: Farber and Bazerman essentially conduct what is now called an (unpaid) "artefactual" field experiment by confronting professional labor arbitrators with hypothetical cases and asking them to choose a settlement under different arbitration schemes. Since then, economists have performed a multitude of experiments evaluating different arbitration schemes, most of them "traditional" laboratory experiments, such as Ashenfelter et al. (1992).

A key question addressed in early experimental studies was the differential effects of conventional arbitration (where the arbitrator can impose any settlement he feels is "fair") versus final-offer arbitration (where the arbitrator is constrained to choose *either* the firm's or the union's final offer). In line with its proponents' expectations, early laboratory experiments (e.g. Notz and Starke, 1978; Starke and Notz, 1981; Neale and Bazerman, 1983) tended to find that final-offer arbitration encouraged more pre-arbitration settlement. These results were strongly rejected by Ashenfelter et al. (1992), who—unlike earlier treatments—provide bargainers with information about past arbitrator decisions in the same scenario (in their experiment, a list of random draws from the same distribution). Since then, Ashenfelter et al.'s main experimental results have been confirmed by Dickinson (2004, 2005) and Deck and Farmer (2003). Deck and Farmer (2007) and Deck et al. (2007) further show that this result is robust to uncertainty over the final value of the quantity bargained over. Also to the credit of conventional arbitration, Kritikos (2006) finds that final-offer arbitration pushes parties to an equal split of the pie 80% of the time, which he interprets as an unnatural distortion of bargaining.

In addition to conventional versus final-offer arbitration, laboratory experiments have been used to study the effects of the following dispute-resolution mechanisms: tri-offer arbitration (an amendment to final-offer arbitration where the arbitrator can also choose a third option provided by an outside "fact finder"); combined arbitration (which imposes final offer arbitration when the arbitrator's "fair" settlement lies between the parties' final offers, and conventional arbitration otherwise); and other schemes known as double-offer arbitration, amended-final-offer arbitration, and automated negotiation. Providing a full description of these results is beyond the scope of the current paper. A recent review, however, is available in Kuhn (2009). By pointing the reader towards this large literature, we hope instead simply to provide an example of a "success story" for laboratory experiments in labor economics: The use of the lab as a testing ground for dispute-resolution mechanisms in labor markets has yielded results of interest not only to readers of *Econometrica*, but also to policymakers who are faced with the real problem of designing dispute-resolution mechanisms, especially in public-sector labor markets

where strikes are prohibited by law. This is clearly a case where "lab labor" has already passed the market test.

5.2. Search

Labor economists have been using sequential search models to interpret field data on unemployment durations since at least Lippman and McCall's classic survey (1976a; 1976b). Search theory also plays an important motivational and interpretive role in field experiments on the determinants of unemployment, such as those conducted on unemployment insurance recipients in Illinois and other states. These studies randomly exposed job searchers to re-employment bonuses and to job search assistance programs, and assessed workers' responses in light of job search theory. Such field experiments have produced useful results regarding the likely effects on workers' job search behavior and other outcomes of making very specific policy changes to existing UI programs. Excellent reviews are available in Meyer (1995, 1996).[84]

This said, both the field experiments and econometric studies of job search and unemployment durations discussed above are probably better suited to evaluating the effects of specific policy changes in specific markets than to testing the most basic predictions of sequential job-search models; indeed many of the estimated effects of policy changes are just as consistent with simple labor-supply models (e.g. Moffitt and Nicholson, 1982) as they are with the search paradigm. The history of laboratory experiments on search, however, focuses much more on the decision processes of individual searchers, and on assessing whether these are consistent with the rules derived from search-theoretic optimization. Given the significant role still played by search models in interpreting microeconomic studies of unemployment and other durations, and in general equilibrium and macroeconomic models of the labor market, laboratory search experiments make an important, but distinct contribution to the economics of labor market search.

To our knowledge the first economists to conduct laboratory search experiments were Schotter and Braunstein (1981, 1982).[85] Subjects engaged in search for a wage drawn at random from a known distribution (either rectangular or triangular), where each draw from the distribution cost the subject c. In the baseline treatment, subjects could take as many draws as they wished, and could accept any of the offers they had received (i.e. there was unlimited recall of past offers). Optimal search theory has strong and well-known predictions in this situation—including the fact that the optimal strategy is a constant reservation wage—and the authors were interested in comparing subjects' actual behavior to these predictions.

[84] For a more recent example of field experiments on job search and UI, see Dolton and O'Neill (2002).

[85] As in the case of bargaining, psychologists conducted laboratory experiments on search long before economists did. See for example Rapoport and Tversky (1966, 1970) and Kahan et al. (1967).

In many ways, the subjects' behavior conformed to theory quite closely. The authors elicited reservation wages by asking workers to bid for the right to search; these reservation wages were, on average, very close to the theoretical optimum for a risk-neutral searcher. Actual search behavior conformed to these elicited reservation wages. Subjects responded to increased search costs and to a mean-preserving spread of the wage distribution as predicted. This said, two interesting anomalies were identified. First, in violation of the theoretical predictions, reservation wages were not constant over time—they fell. (Since reservation wages were optimal at the start of a search spell, this meant that, on average, subjects did not search long enough.) Second, when the subjects were asked to describe their search strategy, most subjects reported that their strategy had two components: (a) a reservation wage, and (b) a maximum number of searches they would conduct. It is tempting to conclude that the latter aspect of the subjects' self-reported strategies explains the falling reservation wage in Schotter and Braunstein's results.

Hey (1982) conducted search experiments similar to Schotter and Braunstein's, (infinite horizon, full recall, constant cost per offer) though in Hey's case the context was consumer search, and subjects did not know what distribution the offers were drawn from (it was normal in practice). Despite this, over 40% of subjects stopped searching with the first price below the theoretically optimal reservation price. Hey's study is however most interesting for the methodological innovation (in economics) of *protocol analysis* in which the subjects are asked to report their thinking during the session out loud, and their words are recorded. It is interesting to compare these self-reported strategies to the predictions of theory. Interestingly, while only a small minority reported a pure reservation-wage strategy, many of the rule-of-thumb strategies described by the subjects performed well, both as (sometimes imperfectly) implemented by the agents themselves *and* as mechanically played by a computer against the price distribution in the experiment. Hey's paper thus illustrates another potential advantage of laboratory experiments—the ability to acquire information about strategies more directly than inferring them from choice behavior alone, which can often present serious econometric challenges in both the field and the lab.

During the next decade, a number of laboratory experiments explored the predictions of search theory under various scenarios, including a finite horizon (Cox and Oaxaca, 1989, 1992), variable search intensity (Harrison and Morgan, 1990), and search from unknown wage distributions (Cox and Oaxaca, 2000). Like most of these studies, Sonnemans (1998) finds that, while overall search efficiency is high (subjects stopped optimally in more than 60% of cases), subjects stop searching too early on average; risk aversion cannot fully account for this. Sonnemans elicits information about subjects' strategies in two ways (apart from their search behavior): one is subjects' use of an electronic information board, which gave subjects information about various aspects of their own past experience, including the number of draws, the highest bid to date, total costs and earnings to date, etc. It seems reasonable to infer that if a subject consulted

this information, it was an input into her strategy. Second, after conducting a sequential search, subjects were asked to conduct another search using the strategy method: they gave explicit instructions to the computer, which conducted the search on their behalf.[86] Interestingly, only 22% of such strategies were of the pure reservation wage form, and only a minority of these picked an optimal wage close to the optimal level. A significant share of such strategies had a satisficing property of stopping when total earnings exceeded a threshold.

More recent laboratory experiments on search include Carbone and Hey (2004), who study the excess sensitivity of consumption to income often noted in macroeconomic data; they find that subjects also over-react when facing similar problems in the lab and argue that agents' limited ability to plan ahead may help explain this phenomenon. Gabaix et al. (2006) explore the ability of a specific model of limited rationality—directed cognition—to account for laboratory evidence on search problems, finding that the directed cognition model better matches the lab evidence whenever its predictions diverge from full rationality. Schunk (2009) considers instead the ability of reference point updating (in a model where subjects are assumed to care about their total net earnings—including those that are sunk—from the search) and loss-aversion to explain behavior in a laboratory search problem. As in many previous studies, on average Schunk's subjects end their search too soon; he argues that a model in which a subset of agents use suboptimal strategies that depend in part on sunk costs accounts well for this phenomenon. Additional results in Schunk and Winter (2009) support the claim that "early stopping" is related to loss aversion, rather than, for example, simple risk aversion.

The most recent paper to address the "early stopping"/declining reservation wage phenomenon in the lab is Brown, Flinn and Schotter (forthcoming). Like other authors, they find sharply declining reservation wages over time.[87] By comparing subjects' behavior in a treatment where offers arrive at a fixed rate but with stochastic cost to one where the time between offers is random but there are no search costs, the authors are able to distinguish hypotheses (such as Schunk and Winter's loss-aversion model) in which searchers respond to accumulating search costs from a competing scenario in which the subjective costs of searching rise sharply over time. They find that rising subjective search costs play a more important role, and argue that the behavioral factors they identify in the lab may help explain the phenomenon of declining reservation wages and/or increasing re-employment hazards in field data. Alternatively, laboratory subjects' rapidly rising time costs may simply reflect some students' short-term scheduling commitments outside the economics laboratory that are not relevant to real-world job searches.

In sum, laboratory experiments on search have, to date, focused almost exclusively on the "partial equilibrium" question of choices by individuals facing an exogenously given

[86] See also Sonnemans (2000) for a more direct comparison of search decisions under the strategy method versus decisions made "by hand".

[87] Brown et al. elicit subjects' reservation wages directly in advance of *every* offer by asking the subjects to name such a wage; a computer then automatically accepts all offers above this and declines the rest.

wage distribution.[88] In doing so, laboratory experiments have exploited the strengths of that methodology, which include the ability to compare behavior to the exact numerical predictions of a theoretical model, and the ability to study the searchers' *strategies* in a dynamic choice situation more directly than attempting to infer strategies from choice histories alone. The main findings of the literature are that, while (especially early in the search process) many subjects behave *as if* they pursued a theoretically optimal strategy, subjects tend to search too little overall, with falling reservation wages and rising acceptance hazards over time. In addition, a majority of subjects report that their strategies depended on factors other than the highest offer received (including total earnings in the search), and indeed selected such strategies when required to play the game using the strategy method. Playing such strategies did not cost subjects much in the search situations examined in the lab, but this does not imply that such strategies have low costs in other environments, including some that are encountered in the real world. Ongoing research continues to attempt to understand these "anomalies" in the light of different behavioral models. What remains unclear is to what extent these anomalies help shed light on job search in real labor markets. Still, the notion that, even under ideal laboratory conditions, subjects systematically choose search strategies that are suboptimal in a particular way, raises interesting hypotheses for research using field data, and for structural models of job search.

5.3. Labor markets

It is probably fair to say that the vast majority of existing labor market experiments focus on interactions between a single "firm" and "worker", whose payoffs if they choose not to interact are exogenously set by the experimenter. That said, as mentioned in Section 2, experimenters have sometimes implemented various forms of markets in which workers compete for jobs (and/or firms compete for workers) in the lab. These papers fall into two main categories, the first of which—the "market design" papers—studies the effects of different institutional mechanisms on labor market performance. The focus here is typically on thin markets where match quality matters a lot, as in the markets for professional workers. Worker effort decisions are typically not considered. The second group of papers is primarily interested in the effects of contractual incompleteness regarding worker effort on the functioning of labor markets.

5.3.1. Market design

One highly successful set of laboratory experiments on labor markets forms part of a larger literature that is sometimes referred to as "market design" (for a recent survey and assessment, see Roth (2008)). This literature draws on a combination of economic theory, market case studies, computation, field experiments and lab experiments to compare the performance of different forms of decentralized market mechanisms and centralized

[88] For an exception in the context of consumer search, see Abrams et al. (2000).

clearinghouses (the latter often designed by the researcher) for specific goods. Markets that have been studied in this literature include airport landing rights (Grether et al., 1981; Rassenti et al., 1982); college admissions (Roth and Sotomayor, 1989); sorority rushes (Mongell and Roth, 1991); macroeconomic risks (Shiller, 1993); newly privatized firms in transition economies (Svejnar and Singer, 1994); postseason college football bowls (Roth and Xin, 1994); radio spectrum licenses (McMillan, 1994); space shuttle payload priorities (Ledyard et al., 2000); student housing (Chen and Sonmez, 2002); electric power (Wilson, 2002); internet auctions (Ariely et al., 2005); human kidneys (Roth et al., 2007); and admissions to New York City high schools (Abdulkadiroglu et al., 2009). Laboratory experiments play an important role in this literature, both in testing hypotheses about particular market mechanisms and in comparing the mechanisms' performance; one lab experiment in this area that may be of particular interest to labor economists is Chen and Sonmez's (2006) study of school choice mechanisms.

Labor markets that have been studied in the "market design" literature include those for American physicians (Roth, 1984, 1986; Roth and Peranson, 1999); baseball players (Nalbantian and Schotter, 1995); clinical psychologists (Roth and Xing, 1997); British physicians (Kagel and Roth, 2000); gastroenterologists (Niederle and Roth, 2003; McKinney et al., 2005); and law clerks (Haruvy et al., 2006; Avery et al., 2007). Of these labor market studies, Nalbantian and Schotter (1995), Kagel and Roth (2000), McKinney et al. (2005) and Haruvy et al. (2006) all use laboratory experiments as at least part of their research design.[89] Key issues studied in these investigations include the causes of "unraveling" that is observed in some markets (e.g. for law clerks) where offers are made so early that little information is available about candidates, and the design of mechanisms that effectively ensure enough market thickness to allow for efficient matching.

Lab experiments play a number of roles in the above studies of labor markets, one of which is to isolate the effects of different institutional changes that cannot be identified in field data, because they co-vary in the field (for example, Niederle and Roth (2009) use the lab to disentangle the effects of exploding offers and binding acceptances). While this could in principle be done in a field experiment, the time, expense, and ethical barriers involved place strict limits on the amount of experimental manipulation that can be done. In contrast, after cross-validating the lab and field in a baseline case, one can quickly and cheaply compare many design permutations in the lab. Market designers are also interested in the robustness of market mechanisms to large potential swings in demand and supply (McKinney et al., 2005); in most cases these cannot be practically (or ethically) manipulated in a field experiment. A final role of lab experiments is simply to provide a level of detail on how individual actions change, for example during transition between market regimes, that is unavailable in field data (Kagel and Roth, 2000).

[89] Recent examples of laboratory experiments with applications to two-sided matching markets include Niederle and Roth (2009) and Niederle et al. (2009).

In sum, the literature on market design, like the arbitration literature, is another economics "success story", where economic theory combined with a carefully chosen set of research methods *including* lab experiments has created institutions that improve the process of economic exchange in the real world.[90] This includes some labor markets, especially for professionals—where markets are thin and match quality is arguably highly heterogeneous—and some allocation problems in education (such as school admissions) that are of considerable interest to labor economists.

5.3.2. Contractual incompleteness and labor markets

While the vast majority of laboratory experiments on principal-agent interactions incorporate labor markets in only the most minimal of senses (as noted, the experimenter typically fixes the two parties' outside options) it is noteworthy that one of the earliest papers to demonstrate the importance of social preferences in principal-agent interactions (Fehr et al., (FKR) 1993) included an explicit labor market, and was explicitly interested in the effect of social preferences on the nature of labor market equilibrium. As already noted, this experiment modeled the labor market as a one-sided auction in which firms posted contract offers to all workers in the session. A main result was that these markets failed to clear when effort was not contractible: firms posted above-market-clearing wages because these wages elicited more effort than market-clearing wages. Firms that tried to underbid such wages did poorly due to workers' endogenous effort responses; involuntary unemployment resulted. This result is evocative of Shapiro and Stiglitz's (1984) well known model of involuntary unemployment, though the mechanism (reciprocity by workers) is different from the disciplinary dismissals that are central to Shapiro and Stiglitz's model.

Since FKR 1993, the most influential paper on the effects of contractual incompleteness on the nature of market interactions is probably Brown et al. (BFF) (2004). This experimental paper is framed as a set of repeated interactions between firms and workers and generates, in our view, a set of results that may significantly influence how labor economists conceptualize labor markets. In BFF's experiments, ten workers interacted with seven firms over fifteen trading periods. Each trading period had two stages: a market for contracts, followed by the exchange of effort for pay in contracts that had been concluded. In all cases, the market for contracts was, again, a one-sided auction: firms posted offers (consisting of a wage, a desired effort level, and the firm's ID number), which could be accepted or rejected by workers. Once a firm's offer was accepted by a worker, both the firm and worker were removed from the market for that trading period. Importantly, firms could make *both private and public offers*: private offers were only conveyed to the worker with whom the firm wanted to trade.

Under the above conditions, BFF compared two main experimental conditions: under complete contracts (C), the firm's required effort level was automatically

[90] See Roth (2002) for a review of methodological issues in the market design literature.

implemented (by the computer) if a worker accepted a contract. Under incomplete contracts (ICF), this third party enforcement of desired effort levels was absent.[91] In both treatments, firms only observe the (current and past) effort levels of their "own" workers. Perhaps unsurprisingly, BFF find that under complete contracts, markets resemble the textbook case: contract offers are public, long-term relations between firms and workers are absent, trading parties seem to be indifferent to their partners' identities, and rent sharing is driven toward the competitive equilibrium (where firms retain all the surplus). Under incomplete contracts, successful exchange is usually initiated by a generous contract offer that a worker reciprocates with a high effort level. After that, firms repeatedly seek out the same worker with a private offer. Rents are shared in these long-term bilateral relationships, which are disciplined by the threat of non-renewal. The result, rather than a classically competitive market, is a market where bilateral monopoly emerges endogenously once the employment relationship has been established, a change described eloquently by Williamson (1985) as the "fundamental transformation".

In many ways, the labor markets that emerge under BFF's ICF condition are much more familiar to today's workers than a "classical" labor market: Workers invest in their reputations early in their careers by working hard; this reputation is (rationally) rewarded by their employer by continuing employment at a "good" wage; both parties' actions are disciplined by the (rational) fear that their contract will not be renewed if they do not continue to cooperate by providing 'fair' levels of wages and effort. For the latter reason, BFF's results are thus more directly supportive (than FKR 1993) of Shapiro and Stiglitz's (1984) model where disciplinary dismissals generate unemployment. In sum, BFF (2004) show very elegantly that if contractual incompleteness plays an important role in firm-worker interactions, the correct theoretical model of labor market equilibrium may not be the competitive one we are used to, nor one of the several interesting equilibrium search models that have been developed in the past two decades (e.g., Mortensen and Pissaridies, 1994), but one in which reciprocity and bilateral monopoly play important roles.

A number of interesting variations on BFF's (2004) design have yielded additional insights. Perhaps the most obvious such variation is replacing the excess supply of workers by an excess supply of firms (Brown et al., 2008). Interestingly, long-term employment relationships still emerge, in which firms pay workers above the going market rate and workers reciprocate with higher effort. Market performance remains high, suggesting that unemployment is not required to enforce high effort levels, though long-term relationships are less frequent than in the excess-worker case. Falk et al. (2008) consider the effects of two institutions—dismissal barriers and bonus pay—in the BFF (2004)

[91] This condition is labeled ICF because workers' ID numbers were fixed throughout the entire session (as they were in the C condition). This allowed firms to target their offers to specific workers (presumably to a worker who had provided high effort in the past), and distinguishes it from an "ICR" treatment where contracts were incomplete but workers' identities could not be tracked during a session.

model. Perhaps unsurprisingly, given the important disciplining role of contract renewal risk, legislated dismissal barriers lead to large reductions in equilibrium effort and market efficiency. Giving firms the option to reward high effort with bonus pay, however, offsets much of these inefficiencies; interestingly, the resulting equilibrium exhibits many fewer long-term relationships.[92]

So far, all the papers in this subsection have considered the effects of hidden *actions* (effort choices) on the functioning of labor markets. We therefore conclude by noting a number of papers that have considered the effects of hidden information (adverse selection) on market equilibrium in the laboratory as well. Interestingly, experimental studies of signaling and screening models have quite a long history, dating back at least to Miller and Plott (1985). These papers may not be familiar to labor economists since they are typically cast in the context of markets for insurance (especially in the screening case), or product quality (signaling). For a recent experimental study of both signaling and screening that is cast in a labor market context, and that reviews the earlier experimental literature, see Kübler et al. (2008). Other recent experiments that have considered the effects of hidden information (about worker ability) in labor markets are Cabrales et al. (2006) and Charness et al. (2008).

5.4. Gender, race and discrimination: Insights from the lab

The existence of discrimination and its effect on employment and wages is a major topic in labor economics. Discrimination is typically considered with respect to gender or ethnic considerations, but may also be relevant in relation to attractiveness.[93] A fundamental issue is how to ameliorate concerns such as the male-female wage gap and racial disadvantage. Can policies such as affirmative action be effective or are these likely to be counter-productive? Once again, it is possible to focus specifically on these issues using laboratory techniques. We review the relevant economics literature in turn.[94]

[92] Two other papers that insert new variations into repeated experimental labor markets similar to BFF(2004) are Healy (2007) and Schram et al. (2007). Healy considers the ability of group reputations to substitute for individual reputations; Schram, Brandts and Gerxhani ask what happens when firms can see workers' past effort levels at *other* firms when bilateral contracts are negotiated.

[93] Discrimination may be based on bias or due to accurate perceptions about differences in ability or other characteristics across groups. While we do not discuss this at length in this section, there is experimental evidence in this regard. For example, Fryer et al. (2005) have "green" and purple" workers with different "investment costs". Investment improves the chance the worker does well on a pre-employment test; this is observed by the firm, who can hire either green (lower cost) or purple. Green workers were hired substantially more often.

[94] There is a substantial and active laboratory literature on racial, gender and other forms of discrimination in the fields of psychology and management science; adequately summarizing this literature would go far beyond the bounds of the current review. A common research design in these papers has laboratory subjects (who in many cases are employed as recruiters in "real life") evaluating job candidates whose race or sex is exogenously manipulated by the experimenter (see Cohen and Bunker, 1975 for a highly-cited example regarding sex, and Dovidio and Gaertner, 2000 for a more recent example regarding race). Other interesting approaches include scenario-based experiments such as Levi and Fried's (2008) on affirmative action, and internet experiments where subjects play on-line games, interacting with avatars of different races (race is manipulated via the other player's first name and the skin tone of the avatar) (Goodwin et al., 2010).

5.4.1. Gender

One of the puzzles in labor economics is the gender-wage gap. Even though the provision of equal opportunities for men and women has been a priority for decades, large gender differences prevail in competitive high-ranking positions. Why do women receive less pay than men? Some of this may stem from a difference in tastes for negotiation. For example, Babcock and Laschever (2003) suggest that women are more likely to shrink from negotiation, potentially costing themselves thousands of dollars in not asking for promotions or in choosing career paths that don't involve negotiation. Another factor may be a difference in risk aversion: Charness and Gneezy (2009) find a very consistent result across a number of experimental studies in that women are more financially risk averse than men. Related to this point, Niederle and Yestrumskas (2008) offer participants a choice of a hard or an easy task, finding gender differences in seeking challenges: Men choose the hard task about 50% more frequently than women, independent of performance level.

Stereotypical attitudes towards female workers may well be at the heart of discriminatory gender-based compensation. Schwieren (2003) considers the issue of why women receive lower pay for comparable work in a gift-exchange experiment in which each firm knows whether workers are male or female. She adapts the double-auction format from Fehr and Falk (1999), with 4 firms and 6 workers in each session. Each group of workers in a session was all-female or all-male, as was each group of firms in a session; she varied whether male workers were paired with male or female workers and *vice versa*. The results are striking, in that female workers receive significantly lower wages than male workers, *even when women are in the role of the firm*. But this doesn't pay for firms, as a high discrepancy between the wage requested and the wage offered leads to low effort. The results suggest that stereotypes are the driving force, rather than statistical discrimination. Women also learn that they cannot successfully ask for high wages and reduce their bids over the course of a session.

The main thrust of experimental research on gender differences in performance is based on the notion that women are less inclined to compete. The first experiment to demonstrate this point is Gneezy et al. (2003), who find that women appear to be less effective than men in competitive environments, despite the fact that their performance is similar to that of males when the environment is not competitive. The experimental task was solving mazes, with six people in each session. In one treatment, people were paid two units for each maze solved; in a second treatment, where there were three men and three women in a session, the person who solved the most mazes received 12 units for each solved maze; the final treatment was identical to the second, except that there were either six men or six women in each session.[95] There is a definite performance gap across gender in tournaments, even though there is no significant

[95] An additional treatment in which one person was chosen at random to receive 12 units per solved maze led to results quite similar to these in the first treatment.

difference in performance across gender in the non-competitive treatment. This is driven by an increase in men's performance in the tournament, as women's performance does not differ across the tournament and piece-rate cases. Furthermore, the effect is more pronounced when women compete against men than when women compete against women, as men's tournament performance does not differ according to gender composition, while women's tournament performance in the same-sex treatment is significantly higher than in the piece-rate condition.

This result is confirmed in Gneezy and Rustichini (2004), a field experiment with school children (aged 9-10) in Israel. The performance measure was the time needed to run 40 meters. Each child first ran alone; the teacher then matched the students in pairs, with the two fastest children paired, then the next two fastest children paired, etc. The children in each pair then ran along side each other. The times in the first task were very similar for boys and girls; however, there were significant gender differences in the change in times from the first run to the second. When girls ran with other girls, the time actually increased slightly; however, the opposite effect was observed when boys ran with other girls. In mixed pairs, the reduction in time for boys was much larger than the reduction in time for girls.

Niederle and Vesterlund (2007) provide evidence that women "shy away" from competition in a task involving adding up sets of five two-digit numbers. In stage 1, participants performed this task under a non-competitive piece-rate scheme, and then performed the task in a four-person tournament in stage 2. In stage 3, people then chose whether they wished to choose the piece-rate scheme or the tournament (in the latter case, they competed against the performance of the other group members in stage 2). There are no significant differences in performance across gender, yet men select the tournament in stage 3 more than twice as frequently (73% versus 35%) as do women. This difference in rates remains significant even when differences in risk attitudes are taken into account. The authors conclude: "...the tournament entry gap is driven by men being more overconfident and by gender differences in preferences for performing in a competition. The result is that women shy away from competition and men embrace it." (p. 1067).

Gneezy et al. (2009) perform field experiments (involving tossing a tennis ball into a bucket) in Tanzania and India. People chose whether they wished to be paid by a piece-rate scheme or to participate in a two-person tournament. The Maasai society in Tanzania is a patriarchal society while the Khasi society in India is matrilineal. The results in the patriarchal society correspond closely to the results in Western cultures, as the Maasai men choose to compete about twice as frequently as do the Maasai women. However, it is quite interesting that the comparison across gender goes in the opposite direction in the matrilineal society. Their results strongly suggest that it may be differences in culture, rather than inherent genetic traits, that drive the results observed in previous experiments.

Two other experiments consider the context of the environment in relation to gender differences in risk-taking and in choosing to compete. Booth and Nolen (2009a) conduct an experiment in which participants choose between a risky gamble and a certain one, with the expected value of the risky gamble being higher than that of the certain outcome. They compare girls who attended coed or single-sex schools. The main finding is that girls from single-sex schools are substantially more likely to choose the gamble than girls from coed schools; in fact, girls from single-sex schools choose the gamble roughly as frequently as do boys. This indicates that social learning and nurture may well be the source of the gender differences found in previous studies. Booth and Nolen (2009b) consider whether girls from single-sex schools differ from girls from coed schools in terms of willingness to compete. They find substantial differences, with girls from single-sex schools more likely to choose competition and behaving more like boys with respect to competitiveness.

5.4.2. Minorities

Discrimination occurs not only with respect to gender, but also with respect to ethnic background. As it is considerably more difficult to arrange experiments in which ethnic background is systematically varied, there are fewer studies in this area.

Fershtman and Gneezy (2001) consider discrimination in Israeli society between Ashkenazic Jews and Eastern Jews; the former tend to be wealthier and better educated. Three different experimental games were used in their study. First, in the investment ("trust") game, three times as much (17.16 versus 5.62) was sent to Ashkenazic male responders as to Eastern male responders; this pattern was observed for both male Ashkenazic and Eastern first-movers (in fact, this discrimination was only observed for males). However, there was no significant difference in the amounts returned by Ashkenazic and Eastern responders. Since first-mover behavior could reflect preferences *per se* rather than trust, a dictator-game control was also conducted, with the amount transferred to the recipient tripled, as in the investment game. Here there is no difference (5.6 versus 5.1) in the amounts sent to male Ashkenazic and Eastern recipients by male dictators, indicating that there is no *per se* taste for discrimination, but rather that the investment game results were driven by (mistaken) beliefs about the expected returns from each ethnic group. Finally, an ultimatum game was conducted to test for the stereotype that Eastern Jews are driven more by a sense of "honor" and so would be more likely to reject small proposals. In fact, Eastern male responders received significantly higher proposals (8.4 versus 5.9) than did Ashkenazic male responders, even though rejection rates for lopsided proposals were nearly the same for both groups. Overall, the observed discrimination in this paper appears to reflect stereotypes rather than rational statistical discrimination.

Fershtman et al. (2005) attempt to distinguish between discrimination against people and "nepotism" in favor of one's own group. At issue is the notion of anonymity rules, which forbid disclosure of group affiliation or background. Using the investment

game, they compare behavior when group identity is fully observable and when it is not. The paper considers two different societies: Belgian (Walloons versus Flemish) and Israeli (ultra-orthodox religious versus secular). Belgian society shows evidence of discrimination, as both Walloons and Flemish sent much more to the responder when he or she was of the same background as the first mover than when the responder was of the other background (1200 versus 745 and 1009 versus 536, respectively). On the other hand, transfers made to anonymous responders were about the same as transfers to own-group responders for both Walloon and Flemish first-movers. However, a very different pattern was observed in Israeli society. In this case, ultra-orthodox religious Jews favor members of their own group, but send similar amounts to both secular Jews and anonymous responders.

5.4.3. Beauty

Does one's level of attractiveness matter in terms of hiring practices and behavior in the workplace? Many people feel that good-looking people are highly advantaged and that looks are essential for being hired in some occupations (e.g. pharmaceutical representatives selling to physicians). Job applicants are routinely counseled to look their best for job interviews. Some recent experimental studies have tested this issue in the laboratory.

Mobius and Rosenblat (2006) investigate the "beauty premium" in an experimental labor market in which firms choose wages for workers in a real-effort task involving solving mazes. Physical attractiveness is found to be uncorrelated with skill in this task. The experimental design varies the degree of oral and visual interaction between firms and workers (and also elicits worker confidence). In the baseline, the firm only observes the "resume" of the worker, which includes the labor-market characteristics of the worker; in the visual treatment, the firm also sees a photograph of the worker; in the oral treatment, the firm sees the resume and also has a telephone conversation with the worker; in the face-to-face treatment, this conversation is face-to-face. They find a substantial beauty premium, which they decompose into three transmission channels: Physically attractive workers are more confident and higher confidence increases wages, these workers are also (incorrectly) considered to be more capable by firms, and these workers also have better oral skills that raise their wages.

Wilson and Eckel (2006) also investigate the beauty premium, using the investment game and photographs. They find a modest beauty premium, in that the average amount sent to an "unattractive" ("attractive") counterpart was $4.64 ($4.98); attractive trustees appear to reciprocate this trust. On the other hand, there is also a "beauty penalty" for attractive first movers, as responders return 35% (30%) of the amounts sent by "unattractive" ("attractive") first movers. It appears that this is driven by disappointed expectations, as responders withhold repayment in this case. Furthermore, responders expect more from attractive first movers, so their expectations are unmet more frequently

and so this withholding is exacerbated when the first mover who disappointed the responder is attractive.

Andreoni and Petrie (2008) study the beauty premium in a public-goods game. When people are not told the contributions of others, there is indeed a beauty premium: attractive people earn 7% more than people of intermediate attractiveness and 12% more than unattractive people. In another treatment, participants are told the individual contributions of each other player. In this case, the beauty premium becomes a beauty penalty, as it appears (as in Wilson and Eckel) that people expect (hope for?) more from attractive people. As these expectations are more likely to be unmet, attractive people seem relatively more selfish, leading to decreased contributions from other participants. There are also some gender effects, as men make larger contributions and this appears to "lead" other people to contribute more in subsequent periods.

5.4.4. Affirmative action

A thorny policy issue is that of affirmative action, whereby some group that is considered to be generally disadvantaged is given an immediate advantage. Emotions run high on this topic, and "reverse discrimination" lawsuits are not uncommon. What are the effects of affirmative action? Clearly this helps disadvantaged people (at least in the short run), but it may well be detrimental with respect to overall performance. Some recent laboratory experiments have investigated this issue in controlled settings.

Schotter and Weigelt (1992) find, in a two-person tournament setting, that the effects of affirmative action depend on the severity of the cost disadvantage for the non-minority group. When the tournament was "unfair", one of these people had to win by a certain amount and/or was assigned a cost function greater than that for the other person. Affirmative action is found to discourage disadvantaged agents from "dropping out", as was observed without affirmative action. Nevertheless, overall the effect of affirmative action on output appears to depend on the degree of disadvantage. When this is large, then affirmative-action programs did improve profits for the "tournament administrators"; however, this did not occur when the cost disadvantage was less severe.

Corns and Schotter (1999) provide an "existence proof" that there exist parameters under which price-preference auctions (often used to promote minority representation in government contracting) can increase minority representation while being cost effective; however, there is a decrease in cost effectiveness if the degree of price preference is too large. The experimental design involves high-cost firms (simulating minority firms) that receive either a 0%, 5%, 10%, or 15% price preference. There were six participants in each session, with costs drawn from a uniform distribution of [100, 200] for the four low-cost firms and [110, 220] for the two high-cost firms. The frequency of a high-cost firm winning the auction increased from 12% in the 0% treatment to 43% in the 15% treatment, largely in line with the theoretical predictions. The average observed price was 121.24 in the 0% condition, 119.29 in the 5% condition, 122.84 in the 10% condition, and 124.41 in the 15% condition. There is no significant difference in the distribution of

drawn costs across any two price-preference treatments. Nevertheless, the results indicate that it is possible to use price preferences to increase representation of "minority" firms without increasing the overall cost or the average observed price.

Niederle et al. (2008) follow up the Niederle and Vesterlund (2007) study by examining how affirmative action affects competitive entry and performance. If women shy away from competition, it may be advantageous to institute an affirmative-action policy under which women are favored in competitions. Participants are formed into groups of six, including three men and three women, and add up a series of five two-digit numbers; in one task, they are paid $0.50 for each correct answer, while in another task the two best performers are paid $1.50 for each correct answer. In the third task, people choose either the piece-rate or the tournament payment scheme. Next there is an affirmative-action task, in which the two winners in a tournament are the highest-performing woman and the highest performer of the other people in the group. The results indicate that when women guaranteed equal representation among winners, more women and fewer men enter competitions. Results also suggest that affirmative action need not be costly, but may be sensitive to parameters, as in Corns and Schotter (1999). Measured beliefs about one's rank and attitudes concerning competition are affected by the affirmative-action policy. While affirmative action may result in reverse discrimination towards men, the authors claim: "... this need not be the outcome when competitive entry is not payoff maximizing. The response in entry implies that it may not be necessary to lower the performance requirement for women to achieve a more diverse set of winners". Thus, affirmative-action policies may potentially be beneficial.

Balafoutas and Sutter (2009) investigate three alternative policy interventions that should promote women in competitive environments: *Preferential Treatment*, *Repetition of the Tournament*, and *Affirmative Action*. These environments are compared with respect to the willingness of men and women to enter a tournament (addition exercises) and the impact on both tournament efficiency (selecting the most qualified people to be the winners) and post-tournament efficiency when groups then participate in a minimum-effort coordination game (the minimum effort translates into the efficiency level). Affirmative-action policy interventions lead to significantly higher entrance rates of women, except for in the *Repetition* treatment. Tournament efficiency does not appear to deteriorate when these policies are introduced; in fact, a "small push" (giving women credit for one additional solved exercise) leads to the highest efficiency. Furthermore, efficiency in a post-tournament teamwork-task is slightly higher in the treatments with successful policy interventions; thus, to the extent that this experimental game serves as a proxy for dynamic considerations, cooperation is not harmed by the introduction of affirmative action.

Calsamiglia et al. (2009) consider affirmative action in two-player tournaments involving solving mini-Sudoku puzzles. Elementary school children at two otherwise-identical private schools in Spain were paired, with the person solving more (net)

puzzles winning the tournament. In one of these schools, the students had experience with these Sudoku puzzles. These policies included adding some number of points to the disadvantaged students total and multiplying the number of solved puzzles by disadvantaged students by some factor greater than one. Their results indicate that performance was not reduced for either advantaged or disadvantaged subjects and that it was in fact enhanced. Additionally, while affirmative action balanced the proportion of disadvantaged individuals winning their respective tournament, the average performance of the pool of winners only decreased slightly.

6. CONCLUSIONS

Laboratory experiments have been used to study almost every aspect of labor economics, ranging from the effects of final-offer arbitration in public sector labor disputes to the causes of gender wage differentials. This vast literature includes a number of noteworthy "success stories", including the use of laboratory experiments as testing grounds for proposed institutional innovations—in particular, matching mechanisms for professional workers, and arbitration systems in public sector union bargaining—that are subsequently used to organize the actual pricing and exchange of labor.

We see laboratory experiments as a useful tool among several at the labor economist's disposal; their key strengths include the ability to control conditions and confounding factors affecting a possible causal relationship more tightly than any other method, the ability to replicate results quickly and easily with newly collected data, the unique ease with which they can test the exact quantitative predictions of simple game-theoretic models, the ability to cleanly study behavior in situations where precise theoretical predictions are absent (as in the case of multiple equilibria), their low cost, the ability to more readily study phenomena that are hidden in the field because they are illegal or disapproved of (such as sabotage, discrimination, and spite), and the ability to elicit data relatively easily on such central game-theoretic concepts as beliefs and strategies that are often difficult to infer from observed behavior alone.

Key objections that have been raised to lab experiments in economics include the artificiality of the experimental context in general and the labor "task" specifically, low stakes, short duration, and the effects of experimenter scrutiny. All of these are important issues, which experimenters should (and typically do) consider seriously. As we have argued in detail in this review, however, none of these concerns—with the possible exception of duration—are insurmountable in the lab, and in general these disadvantages need to be weighed against the lab's potential benefits, enumerated above. In many cases a combination of methods, or a hybrid method—such as the artefactual field experiments and framed field experiments described by List (2009)—may be the best approach.

Perhaps the most widespread concern with lab experiments is the representativeness of the laboratory population relative to the field population of interest; this concern seems particularly acute when discussing lab results that are hard to explain without some

recourse to "non-standard" preferences. As we have noted, this issue can be addressed in a number of ways, including conducting lab experiments directly on the field population of interest and studying the selection process itself (for example, selection into competitive versus co-operative work environments, or into situations where altruism might be expected or not) in the lab.

Greater clarity on "representativeness" can, in our opinion, be also achieved if future investigators are more precise about *which* field environment their experiment is intended to represent. Clearly, if the intended field population is highly specific (for example, new registrants for unemployment insurance in Illinois), and the researcher's main interest is in the *magnitude* of the behavioral response to a contemplated policy change affecting this population, the ideal research design is almost certainly what List (2009) calls a 'natural field experiment' on that population. However, when searching for underlying principles of behavior, derived from a specific theoretical model of strategic interaction and/or social preferences that one hopes might apply to a broader field population (for example, all college-educated employees, or even all humans), a lab experiment on a random sample from a broader group may be more useful. In all cases, the research question should guide the choice of method, and the investigator should specify the population on which the experiment is intended to shed light.

Probably the most frequent application of lab experiments in labor economics has been to principal-agent interactions between workers and firms. While many predictions of what one might call "traditional" principal-agent theory have been confirmed in the lab—for example, in most cases effort increases when marginal work incentives are raised, contributions to team production fall with the number of team members, and ratchet effects disappear when *ex post* labor market competition is intensified—lab experiments have also identified a significant number of robust departures from rational, selfish behavior in principal-agent interactions. As already noted, a catch-all phrase for the modifications to traditional preference structures that have been proposed to explain these "anomalies" is "social preferences"; the discovery and mapping of social preferences is probably the most important contribution of lab experiments to principal-agent analysis, and to labor economics more generally.[96]

That said, much remains unknown about social preferences. On the theoretical side, one area in which progress is needed is in the development of social-preference models that successfully apply to large, multi-worker workplaces. Most current social-preference models have been developed to explain behavior in the lab. While social preferences might explain behavior in these situations, it is remains unclear how these models are best generalized to larger firms with many workers and several hierarchical levels. On the empirical side, social preferences seem capable of changing their shape and

[96] Non-selfish preferences are of course not the only possible explanation of departures from predicted equilibria in effort-supply models. Others include simple limits to human cognition (especially in complex, multi-period games), and practical difficulties of coordinating on Nash equilibrium even when a unique one exists.

form considerably between different contexts, as well as between the lab and the field. Empirical work that sheds light on *which* social preferences (including, among others, inequity aversion, reciprocity, guilt, spite, jealousy and simple altruism) come into play *when* will be of great value.

While much remains to be learned about the effects of compensation systems on worker behavior within firms, it is perhaps worth gauging our progress by considering the state of knowledge on this topic some 15 years ago in the words of some of the main contributors (emphasis added):

> *"There is a large and growing interest in the economic theory of the internal workings of firms. However, this literature is based on very little data and limited stylized facts."*
>
> *(Baker et al., 1994a)*

> *"For a time there was considerable excitement about implicit contract theory …That literature soon fell out of favor, but in its place came more refined information economic analyses that viewed wage contracts as optimal responses to asymmetries in information between employees and firms. With this, the logical possibilities for explaining wage behavior grew dramatically. Today's large variety of models and modeling options has put theory well ahead of observation".*
>
> *(Baker et al., 1994b).*

And finally, from an article titled "Internal Labor Markets: Too Many Theories, Too Few Facts":

> *"With the advent of information economics and contract theory, models of internal labor markets—or at least selected features of these markets—have begun to emerge. The objective of these theories is to show that internal-labor-market outcomes can be construed as second-best solutions to contracting problems under incomplete information. …At this point, there is hardly any feature of internal labor markets that cannot be given some logical explanation using the right combination of uncertainty, asymmetric information and opportunism.*
>
> *(Baker and Holmstrom, 1995)*

Thus, just 15 years ago, research on incentives in organizations consisted to a disappointingly large extent of (a) identifying a "paradoxical" feature of some internal labor market (for example, academic tenure, or pay raises based purely on seniority), then (b) crafting a theoretical model showing how that feature was in fact an efficient response to some contracting problem. Most analyses stopped there.

Today, thanks both to a laboratory literature that began in the mid-1990s and a more recent surge in field experiments, we now face a cornucopia (some might say, an unmanageable deluge) of facts about agents' and principals' behavior in the above models. Indeed, after reviewing this literature, it is tempting to conclude that we now have "Too Many Facts, Too Few Theories". Certainly, an important part of the road ahead is the development of models (that can account for the robust departures from the

predictions of 'traditional' principal-agent models) both in the lab and in the field. In our view, a variety of empirical tools including both lab and field experiments, natural experiments, econometric studies, and calibration of structural models will provide the necessary empirical discipline in this next phase of research on one of the most basic issues in labor markets: the exchange of effort for pay within firms.

REFERENCES

Abdulkadiroglu, Atila, Pathak, Parag A., Roth, Alvin E., 2009. Strategy-proofness versus efficiency in matching with indifferences: redesigning the NYC high school match. American Economic Review 99 (5), 1954–1978.

Abeler, Johannes, Falk, Armin, Goette, Lorenz, Huffman, David, 2009. Reference points and effort provision. IZA Discussion Paper No. 3939, Institute for the Study of Labor.

Abeler, Johannes, Altmann, Steffen, Kube, Sebastian, Wibral, Matthias, 2009. Gift exchange and workers' fairness concerns: when equality is unfair. Journal of the European Economic Association (forthcoming).

Abrams, Eric, Sefton, Martin, Yavas, Abdullah, 2000. An experimental comparison of two search models. Economic Theory 16 (3), 735–749.

Acemoglu, Daron, Kremer, Michael, Mian, Atif, 2008. Incentives in markets, firms, and governments. Journal of Law, Economics, and Organization 24 (2), 273–306.

Akerlof, George A., 1976. The economics of caste and of the rat race and other woeful tales. Quarterly Journal of Economics 90 (4), 599–617.

Akerlof, George A., 1982. Labor contracts as partial gift exchange. The Quarterly Journal of Economics 97 (4), 543–569.

Akerlof, G., 1984. Gift exchange and efficiency-wage theory: four views. American Economic Review 74 (2), 79–83.

Akerlof, G., Yellen, J., 1986. Efficiency Wage Models of the Labor Market. Cambridge University Press, New York.

Akerlof, George A., Yellen, Janet L., 1988. Fairness and unemployment. American Economic Review 78 (2), 44–49.

Akerlof, George, Yellen, Janet, 1990. The fair-wage effort hypothesis and unemployment. Quarterly Journal of Economics 105 (2), 255–284.

Altonji, J., Pierret, C., 2001. Employer learning and statistical discrimination. Quarterly Journal of Economics 116 (1), 313–350.

Anderhub, Vital, Königstein, Manfred, Kübler, Dorothea, 2003. Long-term work contracts versus sequential spot markets: experimental evidence on firm-specific investment. Labour Economics 10 (4), 407–425.

Anderhub, Vital, Gächter, Simon, Königstein, Manfred, 2002. Efficient contracting and fair play in a simple principal-agent experiment. Experimental Economics 5 (1), 5–27.

Andreoni, James, Brown, Paul, Vesterlund, Lise, 2002. What makes an allocation fair? Some experimental evidence. Games and Economic Behavior 40 (1), 1–24.

Andreoni, James, Castillo, Marco, Petrie, Ragan, 2003. What do bargainers' preferences look like? Experiments with a convex ultimatum game. American Economic Review 93 (3), 672–685.

Andreoni, James, Petrie, Ragan, 2008. Beauty, gender and stereotypes: evidence from laboratory experiments. Journal of Economic Psychology 29 (1), 73–93.

Ariely, Dan, Ockenfels, Axel, Roth, Alvin E., 2005. An experimental analysis of ending rules in internet auctions. RAND Journal of Economics 36 (4), 890–907.

Ariely, Dan, Gneezy, Uri, Loewenstein, George, Mazar, Nina, 2008. Large stakes and big mistakes. Unpublished paper, Duke University.

Ashenfelter, Orley, Currie, Janet, Farber, Henry S., Spiegel, Matthew, 1992. An experimental comparison of dispute rates in alternative arbitration systems. Econometrica 60 (6), 1407–1433.

Avery, Christopher, Jolls, Christine, Posner, Richard, Roth, Alvin E., 2007. The new market for federal judicial law clerks national bureau of economic research. Inc. NBER Working Papers: 13213.

Babcock, Linda, Loewenstein, George, 1997. Explaining bargaining impasse: the role of self-serving biases. Journal of Economic Perspectives 11 (1), 109–126.
Babcock, Linda, Laschever, Sara, 2003. Women Don't Ask: Negotiation and the Gender Divide. Princeton University Press, Princeton.
Baker, George, Gibbons, Robert, Murphy, Kevin J., 1994. Subjective performance measures in optimal incentive contracts. Quarterly Journal of Economics 109, 1125–1156.
Baker, George, Gibbs, Michael, Holmstrom, Bengt, 1994a. The internal economics of the firm: evidence from personnel data. Quarterly Journal of Economics 109 (4), 881–919.
Baker, George, Gibbs, Michael, Holmstrom, Bengt, 1994b. The wage policy of a firm. Quarterly Journal of Economics 109 (4), 921–955.
Baker, George, Holmstrom, Bengt, 1995. Internal labor markets: too many theories, too few facts. American Economic Review 85 (2), 255–259.
Balafoutas, L., Sutter, M., 2009. Gender, competition and the effects of policy interventions, mimeo.
Bandiera, Oriana, Barankay, Iwan, Rasul, Imran, 2005. Social preferences and the response to incentives: evidence from personnel data. Quarterly Journal of Economics 120 (3), 917–962.
Bandiera, Oriana, Barankay, Iwan, Rasul, Imran, 2007. Incentives for managers and inequality among workers: evidence from a firm-level experiment. Quarterly Journal of Economics 122 (2), 729–773.
Bandiera, Oriana, Barankay, Iwan, Rasul, Imran, 2009a. Social connections and incentives in the workplace: evidence from personnel data. Econometrica 77 (4), 1047–1094.
Bandiera, Oriana, Barankay, Iwan, Rasul, Imran, 2009b. Team incentives: evidence from a field experiment. Unpublished paper, University of Pennsylvania.
Barofsky, I., Hurwitz, D., 1968. Within ratio responding during fixed ratio performance. Psychonomic Science 11 (7), 263–264.
Barr, Abigail, Serneels, Pieter, 2009. Reciprocity in the workplace. Experimental Economics 12 (1), 99–112.
Bartling, B., Fischbacher, U., 2008. Shifting the blame: on delegation and responsibility. IEW Working Paper 380.
Battalio, R.C., Green, L., Kagel, J.H., 1981. Income-leisure tradeoffs of animal workers. American Economic Review 71, 621–632.
Battalio, R.C., Kagel, J.H., 1985. Consumption-leisure tradeoffs of animal workers: effects of increasing and decreasing marginal wage rates in a closed economy experiment. In: Smith, Vernon L. (Ed.), Research in Experimental Economics, vol. 3. JAI Press, Greenwich, Conn., pp. 1–30.
Battigalli, Pierpaolo, Dufwenberg, Martin, 2007. Guilt in games. American Economic Review Papers & Proceedings 97, 170–176.
Battigalli, Pierpaolo, Dufwenberg, Martin, 2009. Dynamic psychological games. Journal of Economic Theory 144, 1–35.
Bauernschuster, Stefan, Dürsch, Peter, Oechssler, Jörg, Vadovic, Radovan, 2009. Mandatory sick pay provision: a labor market experiment. Jena Economic Research Papers in Economics 2009-076, Friedrich-Schiller-University Jena, Max-Planck-Institute of Economics.
Becker, Gary, 1964. Human Capital. National Bureau of Economic Research, New York.
Becker, B.E., Huselid, M.A., 1992. The incentive effects of tournament compensation systems. Administrative Science Quarterly 37, 336–350.
Belot, Michele, van de Ven, Jeroen, 2009. Friendships and favoritism at school—evidence from the field. Unpublished paper, University of Amsterdam.
Bellemare, Charles, Kröger, Sabine, 2007. On representative social capital. European Economic Review 51 (1), 181–202.
Bellemare, Charles, Shearer, Bruce, 2009. Gift giving and worker productivity: evidence from a firm-level experiment. Games and Economic Behavior 67 (1), 233–244.
Berg, Joyce, Dickhaut, John, McCabe, Kevin, 1995. Trust, reciprocity and social history. Games and Economic Behavior 10, 122–142.
Bernhardt, D., Scoones, D., 1993. Promotion, turnover, and preemptive wage offers. American Economic Review 83 (4), 771–791.
Berninghaus, S., et al., 2001. Decentralized versus collective bargaining—an experimental study. International Journal of Game Theory 30 (3), 437–448.

Berninghaus, Siegfried, Werner, Guth, Claudia, Keser, 2003. Unity suggests strength: an experimental study of decentralized and collective bargaining. Labour Economics 10 (4), 465–479.
Bewley, T., 1999. Why Wages Don't Fall During a Recession. Harvard University Press, Cambridge, MA.
Binmore, Ken, Dasgupta, Partha (Eds.), 1987. The Economics of Bargaining. Blackwell, Oxford.
Bénabou, Roland, Tirole, Jean, 2003. Intrinsic and extrinsic motivation. Review of Economic Studies 70 (3), 489–520.
Bénabou, Roland, Tirole, Jean, 2006. Incentives and prosocial behavior. American Economic Review 96 (5), 1652–1678.
Berger, Jonah, Pope, Devin G., 2009. Can losing lead to winning? Unpublished paper, The Wharton School, University of Pennsylvania.
Blount, Sally, 1995. When social outcomes aren't fair: the effect of causal attributions on preferences. Organizational Behavior and Human Decision Processes 63, 131–144.
Bolton, Gary, 1991. A comparative model of bargaining: theory and evidence. American Economic Review 81, 1096–1136.
Bolton, Gary, Ockenfels, Axel, 2000. ERC: a theory of equity, reciprocity and competition. American Economic Review 90, 166–193.
Boning, Brent, Ichniowski, Casey, Shaw, Kathryn, 2007. Opportunity counts: teams and the effectiveness of production incentives. Journal of Labor Economics 25 (4), 613–650.
Booth, Alison L., Nolen, Patrick J., 2009a. Gender differences in risk behavior: does nurture matter? Discussion Paper No. 4026, Institute for the Study of Labor.
Booth, Alison L., Nolen, Patrick J., 2009b. Choosing to compete: how different are girls and boys? Discussion Paper No. 4027, Institute for the Study of Labor.
Bornstein, Gary, Gneezy, Uri, Nagel, Rosmarie, 2002. The effect of intergroup competition on group coordination: an experimental study. Games and Economic Behavior 41 (1), 26–45.
Brandts, Jordi, Charness, Gary, 2003. Truth or consequences: an experiment. Management Science 49, 116–130.
Brandts, Jordi, Charness, Gary, 2004. Do labour market conditions affect gift exchange? Some experimental evidence. Economic Journal 114 (497), 684–708.
Brandts, J., Charness, G., 2009. The strategy versus the direct-response method: a survey of experimental comparisons. Experimental Economics (forthcoming).
Brandts, Jordi, Cooper, David J., 2007. It's what you say, not what you pay: an experimental study of manager-employee relationships in overcoming coordination failure. Journal of the European Economic Association 5 (6), 1223–1268.
Bronars, S., 1986. Strategic Behavior in Tournaments. Texas A&M University.
Brosig, Jeannette, Joachim, Weimann, Chun-Lei, Yang, 2003. The hot versus cold effect in a simple bargaining experiment. Experimental Economics 6 (1), 75–90.
Brosig, Jeannette, Joachim, Weimann, Chun-Lei, Yang, 2004. Communication, reputation, and punishment in sequential bargaining experiments. Journal of Institutional and Theoretical Economics 160 (4), 576–606.
Brown, K., Harlow, W., Starks, L., 1996. Of tournaments and temptations: an analysis of managerial fund incentives in the mutual fund industry. Journal of Finance 51, 85–110.
Brown, Martin, Falk, Armin, Fehr, Ernst, 2004. Relational contracts and the nature of market interactions. Econometrica 72 (3), 747–780.
Brown, Martin, Falk, Armin, Fehr, Ernst, 2008. Competition and relational contracts: the role of unemployment as a disciplinary device. Working Paper No. 3345, Institute for the Study of Labor.
Brown, Meta, Flinn, Chris, Schotter, Andrew, Real time search in the laboratory and the market. American Economic Review (forthcoming).
Bruggen, Alexander, Strobel, Martin, 2007. Real effort versus chosen effort in experiments. Economics Letters 96 (2), 232–236.
Bull, Clive, Schotter, Andrew, Weigelt, Keith, 1987. Tournaments and piece rates: an experimental study. Journal of Political Economy 95 (1), 1–33.
Burks, Stephen, Carpenter, Jeffrey, Goette, Lorenz, 2009. Performance pay worker cooperation: evidence from an artefactual field experiment. Journal of Economic Behavior and Organization 70 (3), 458–469.

Cabrales, Antonio, Charness, Gary, Villeval, Marie-Claire, 2006. Competition, hidden information and efficiency: an experiment. Discussion Paper No. 2296, Institute for the Study of Labor.
Cadsby, C. Bram, Song, Fei, Tapon, Francis, 2007. Sorting and incentive effects of pay for performance: an experimental investigation. Academy of Management Journal 50 (2), 387–405.
Cadsby, C. Bram, Song, Fei, Tapon, Francis, 2008. Are you paying your employees to cheat? An experimental investigation. Working Paper No. 2008-10, University of Guelph.
Cadsby, C. Bram, Song, Fei, Tapon, Francis, 2009. The impact of risk aversion and stress on the incentive effect of performance pay. Unpublished paper, University of Guelph.
Calsamiglia, C., Franke, J., Rey-Biel, P., 2009. The incentive effects of affirmative action in a real-effort tournament, mimeo.
Camerer, C., Hogarth, R., 1999. The effects of financial incentives in economics experiments: a review and capital-labor-production framework. Journal of Risk and Uncertainty 19, 7–42.
Camerer, C., Lovallo, D., 1999. Overconfidence and excess entry. American Economic Review 89, 306–318.
Camerer, Colin, Weber, Roberto, Experimental organizational economics. In: Gibbons, R. Roberts, J. (Eds.), The Handbook of Organizational Economics, Princeton University Press, Princeton, NJ (forthcoming).
Camerer, C., Babcock, L., Lowenstein, G., Thaler, R., 1997. Labor supply of NYC taxi cabdrivers: one day at a time. Quarterly Journal of Economics 112 (2), 407–441.
Cameron, Lisa A., 1999. Raising the stakes in the ultimatum game: experimental evidence from Indonesia. Economic Inquiry 37 (1), 47–59.
Carbone, Enrica, Hey, John D., 2004. The effect of unemployment on consumption: an experimental analysis. Economic Journal 114 (497), 660–683.
Card, David, 1990. Strikes and wages: a test of an asymmetric information model. Quarterly Journal of Economics 105 (3), 625–659.
Carmichael, H. Lorne, 1983a. The agent-agents problem: payment by relative output. Journal of Labor Economics 1 (1), 50–65.
Carmichael, H. Lorne, 1983b. Firm-specific human capital and promotion ladders. Bell Journal of Economics 14 (1), 251–258.
Carmichael, L., MacLeod, W.B., 2000. Worker cooperation and the ratchet effect. Journal of Labor Economics 18, 1–19.
Carmichael, H.Lorne, MacLeod, W. Bentley, 1993. Multiskilling, technical change and the Japanese firm. Economic Journal 103 (416), 142–160.
Carpenter, Jeffrey, Rudisill, McAndrew, 2003. Fairness, escalation, deference, and spite: strategies used in labor-management bargaining experiments with outside options. Labour Economics 10 (4), 427–442.
Carpenter, Jeffrey, Matthews, Peter, Schirm, John, 2010. Tournaments and office politics: evidence from a real effort experiment. American Economic Review 100 (1), 504–517.
Carpenter, Jeffrey, Seki, Erica, Do social preferences increase productivity? Field experimental evidence from fisherman in Toyama Bay. Economic Inquiry (forthcoming).
Charness, Gary, 2004. Attribution and reciprocity in an experimental labor market. Journal of Labor Economics 22 (3), 665–688.
Charness, G., 2010. Laboratory experiments: challenges and promise. Journal of Economic Behavior and Organization 73, 21–23.
Charness, G., Chen, K.-Y., 2002. Minimum advertised-price policy rules and retailer behavior: an experiment. Interfaces 32, 62–73.
Charness, Gary, Corominas-Bosch, Margarida, Frechette, Guillaume R., 2007. Bargaining and network structure: an experiment. Journal of Economic Theory 136 (1), 28–65.
Charness, Gary, Dufwenberg, Martin, 2006. Promises and partnership. Econometrica 74 (6), 1579–1601.
Charness, Gary, Dufwenberg, Martin, 2009. Participation. American Economic Review (forthcoming).
Charness, Gary, Fréchette, Guillaume, Kagel, John, 2004. How robust is laboratory gift exchange? Experimental Economics 7, 189–205.
Charness, G., Gneezy, U., 2009. Strong evidence for gender differences in experimental investing, mimeo.
Charness, Gary, Kuhn, Peter, 2007. Does pay inequality affect worker effort? Experimental evidence. Journal of Labor Economics 25 (4), 693–724.

Charness, Gary, Kuhn, Peter, Villeval, Marie-Claire, 2008. Competition and the ratchet effect. Discussion Paper No. 3784, Institute for the Study of Labor.
Charness, G., Kuhn, P., Villeval, M.C., 2009. Competition and the ratchet effect. Journal of Labor Economics (forthcoming).
Charness, Gary, Levine, David I., 2007. Intention and stochastic outcomes: an experimental study. Economic Journal 117 (522), 1051–1072.
Charness, G., Masclet, D., Villeval, M.-C., 2010. Competitive preferences and status as an incentive: experimental evidence, mimeo.
Charness, Gary, Rabin, Matthew, 2002. Understanding social preferences with simple tests. Quarterly Journal of Economics 117 (3), 817–869.
Charness, Gary, Villeval, Marie-Claire, 2009. Cooperation and competition in intergenerational experiments in the field and laboratory. American Economic Review 99 (3), 956–978.
Charness, Gary, Yang, Chun-Lei, 2008. Endogenous group formation and public goods provision: exclusion, exit, mergers, and redemption. Department of Economics, UC Santa Barbara, Working Paper No. 13-08.
Chaudhuri, Ananish, 1998. The Ratchet principle in a principal with unknown costs: an experimental analysis. Journal of Economic Behavior and Organization 37 (3), 291–304.
Chaudhuri, Ananish, Graziano, Sara, Maitra, Pushkar, Graziano, Sara, Maitra, Pushkar, 2006. Social learning and norms in an experimental public goods game with inter-generational advice. Review of Economic Studies 73 (2), 357–380.
Chaudhuri, Ananish, 2007. Conditional cooperation and social norms in public goods experiments: a survey of the literature. Unpublished Manuscript, University of Auckland.
Chen, Kang, Tang, Fang-Fang, 2009. Cultural differences between Tibetans and ethnic Han Chinese in ultimatum bargaining experiments. European Journal of Political Economy 25 (1), 78–84.
Chen, Yan, Sonmez, Tayfun, 2006. School choice: an experimental study. Journal of Economic Theory 127 (1), 202–231.
Chen, Yan, Sonmez, Tayfun, 2002. Improving efficiency of on-campus housing: an experimental study. American Economic Review 92 (5), 1669–1686.
Chevalier, J.A., Ellison, G.D., 1997. Risk taking by mutual funds as a response to incentives. Journal of Political Economy 105, 1167–1200.
Coffman, L., 2009. Intermediation reduces punishment and reward, mimeo, Harvard University.
Cohen, Stephen L., Bunker, Kerry A., 1975. Subtle effects of sex role stereotypes on recruiters' hiring decisions. Journal of Applied Psychology 60 (5), 566–572.
Cooper, David, Kagel, John, Other regarding preferences: a selective survey of experimental results. In: Kagel, J., Roth, A. (Eds.), The Handbook of Experimental Economics, vol. 2, Princeton University Press, Princeton (in press).
Cooper, David J., Kagel, John H., Lo, Wei, Liang Gu, Qing, 1999. Gaming against managers in incentive systems: experimental results with Chinese students and Chinese managers. American Economic Review 89 (4), 781–804.
Corns, Allan, Schotter, Andrew, 1999. Can affirmative action be cost effective? An experimental examination of price-preference auctions. American Economic Review 89 (1), 291–305.
Cox, James C., Oaxaca, Ronald L., 1989. Laboratory experiments with a finite-horizon job-search model. Journal of Risk and Uncertainty 2 (3), 301–329.
Cox, James C., Oaxaca, Ronald L., 1992. Direct tests of the reservation wage property. Economic Journal 102 (415), 1423–1432.
Cox, James C., Oaxaca, Ronald L., 2000. Good news and bad news: search from unknown wage offer distributions. Experimental Economics 2 (3), 197–225.
Cox, James C., 2004. How to identify trust and reciprocity. Games and Economic Behavior 46, 260–281.
Cox, James C., Deck, Cary, 2005. On the nature of reciprocal motives. Economic Inquiry 43 (3), 623–635.
Cox, James C., Deck, Cary, 2006. When are women more generous than men? Economic Inquiry 44, 587–598.
Cox, J., Friedman, D., Gjerstad, S., 2007. A tractable model of reciprocity and fairness. Games and Economic Behavior 59, 17–45.
Cox, James C., Sadiraj, Klarita, Sadiraj, Vjollca, 2008. Implications of trust, fear, and reciprocity for modeling economic behavior. Experimental Economics 11, 1–24.

Cramton, P.C., Tracy, J.S., 1992. Strikes and holdout in wage bargaining: theory and data. American Economic Review 81, 100–121.
Crawford, Vincent P., 1979. On compulsory-arbitration schemes. Journal of Political Economy 87 (1), 131–159.
Crawford, V., Meng, J., 2008. New York City cabdrivers' labor supply revisited: reference-dependent preferences with rational-expectations targets for hours and income. UCSD Discussion Paper.
Croson, Rachel, Boles, Terry, Murnighan, J. Keith, 2003. Cheap talk in bargaining experiments: lying and threats in ultimatum games. Journal of Economic Behavior and Organization 51 (2), 143–159.
Currie, Janet, McConnell, Sheena, 1991. Collective bargaining in the public sector: the effect of legal structure on dispute costs and wages. American Economic Review 81 (4), 693–718.
Deci, E., Koestner, R., Ryan, R., 1999. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. Psych. Bull. 125, 627–668.
Deck, Cary A., Farmer, Amy, 2003. Bargaining with asymmetric uncertainty: arbitration mechanisms compared. Discussion Paper, University of Arkansas.
Deck, Cary A., Farmer, Amy, 2007. Bargaining over an uncertain value: arbitration mechanisms compared. Journal of Law, Economics & Organization 23 (3), 547–579.
Deck, Cary A., Farmer, Amy, Zeng, Dao-Zhi, 2007. Amended final-offer arbitration over an uncertain value: a comparison with CA and FOA. Experimental Economics 10 (4), 439–454.
Della Vigna, Stefano, List, John A., Malmendier, Ulrike, 2009. Testing for altruism and social pressure in charitable giving. NBER Working Paper No. 15629.
Deutsch, M., Krauss, R., 1960. The effects of threat on interpersonal bargaining. Journal of Abnormal and Social Psychology 61, 223–230.
Dickinson, David L., 1999. An experimental examination of labor supply and work intensities. Journal of Labor Economics 17 (4), 638–670.
Dickinson, David L., 2001. The carrot versus the stick in work team motivation. Experimental Economics 4 (1), 107–124.
Dickinson, David L., 2004. A comparison of conventional, final-offer, and "combined" arbitration for dispute resolution. Industrial & Labor Relations Review 57 (2), 288–301.
Dickinson, David L., 2005. Bargaining outcomes with double-offer arbitration. Experimental Economics 8 (2), 145–166.
Dickinson, David L., Issac, R. Mark, 1998. Absolute and relative awards for individuals in team production. Managerial and Decision Economics 19 (4–5), 299–310.
Dickinson, David L., 2009. The effects of beliefs versus risk attitude on bargaining outcomes. Theory and Decision 66 (1), 69–101.
Dickinson, David L., Villeval, Marie-Claire, 2008. Does monitoring increase work effort? The complementarity between agency and crowding-out theories. Games and Economic Behavior 63 (1), 56–76.
Dohmen, Thomas J., Falk, Armin, 2006. Performance pay and multi-dimensional sorting: productivity, preferences and gender. Discussion Paper No. 2001, Institute for the Study of Labor.
Dolton, Peter, O'Neill, Donal, 2002. The long-run effects of unemployment monitoring and work-search programs: experimental evidence from the United Kingdom. Journal of Labor Economics 20 (2), 381–403 (Part 1).
Dovidio, John F., Gaertner, Samuel L., 2000. Aversive racism and selection decisions. Psychological Science 11 (4), 315–319.
Dufwenberg, Martin, Gneezy, Uri, 2000. Measuring beliefs in an experimental lost wallet game. Games and Economic Behavior 30, 163–182.
Dufwenberg, Martin, Kirchsteiger, Georg, 2004. A theory of sequential reciprocity. Games and Economic Behavior 47, 268–298.
Dürsch, Peter, Oechssler, Jörg, Vadovic, Radovan, 2008. Sick pay provision in experimental labor markets. Discussion Paper No. 476, University of Heidelberg.
Eckel, Catherine C., Grossman, Philip J., 1997. Are women less selfish than men? Evidence from dictator games. Economic Journal.
Eckel, Catherine C., Grossman, Philip J., 2005. Managing diversity by creating team identity. Journal of Economic Behavior and Organization 58 (3), 371–392.

Ederer, Florian P., Fehr, Ernst, 2007. Deception and incentives: how dishonesty undermines effort provision. Discussion Paper No. 3200, Institute for the Study of Labor.
Ellingsen, Tore, Johannesson, Magnus, 2004. Promises, threats and fairness. Economic Journal 114 (495), 397–420.
Ellingsen, Tore, Johannesson, Magnus, 2008. Pride and prejudice: the human side of incentive theory. American Economic Review 98 (3), 990–1008.
Eriksson, Tor, Teyssier, Sabrina, Villeval, Marie-Claire, 2008. Self-selection and the efficiency of tournaments. Economic Inquiry 47, 530–548.
Eriksson, Tor, Poulsen, Anders, Villeval, Marie-Claire, 2008. Feedback and incentives: experimental evidence. Working Paper No. 3440, Institute for the Study of Labor.
Eswaran, Mukesh, Kotwal, Ashok, 1984. The moral hazard of budget-breaking. RAND Journal of Economics 15 (4), 578–581.
Fahr, R., Irlenbusch, B., 2000. Fairness as a constraint on trust in reciprocity: earned property rights in a reciprocal exchange experiment. Economics Letters 66 (3), 275–282.
Falk, A., Fehr, E., 2003. Why labour market experiments? Labour Economics 10, 399–406.
Falk, A., Heckman, J., 2009. Lab experiments are a major source of knowledge in the social sciences. Science 326 (5952), 535–538.
Falk, Armin, Fehr, Ernst, Fischbacher, Urs, 2003a. On the nature of fair behavior. Economic Inquiry 41, 20–26.
Falk, Armin, Fehr, Ernst, Fischbacher, Urs, 2003b. Reasons for conflict: lessons from bargaining experiments. Journal of Institutional and Theoretical Economics 159 (1), 171–187.
Falk, Armin, Ichino, Andrea, 2006. Clean evidence on peer effects. Journal of Labor Economics 24 (1), 39–57.
Falk, Armin, Fischbacher, Urs, 2006. A theory of reciprocity. Games and Economic Behavior 54 (2), 293–315.
Falk, Armin, Zehnder, Christian, 2006. Discrimination and in-group favoritism in a citywide trust experiment. IEW-Working Paper No. 318.
Falk, Armin, Kosfeld, Michael, 2006. The hidden costs of control. American Economic Review 96 (5), 1611–1630.
Falk, Armin, Fehr, Ernst, Huffman, David, 2008. The power and limits of tournament incentives. Unpublished manuscript, University of Bonn.
Falk, Armin, Fehr, Ernst, Zehnder, Christian, 2006. The behavioral effects of minimum wage laws. Quarterly Journal of Economics 121 (4), 1347–1381.
Falk, Armin, Huffman, David, MacLeod, W. Bentley, 2008. Institutions and contract enforcement. Working Paper No. 13961, National Bureau of Economic Research.
Fama, E.F., 1980. Agency problems and the theory of the firm. Journal of Political Economy 88, 288–307.
Farber, Henry S., 1987. The analysis of union behavior. In: Layard, Richard, Ashenfelter, Orley (Eds.), Handbook of Labor Economics, vol. 2. North-Holland, Amsterdam.
Farber, Henry S., 2005. Is tomorrow another day? The labor supply of New York city cabdrivers. Journal of Political Economy 113 (1), 46–82.
Farber, Henry S., 2008. Reference-dependent preferences and labor supply: the case of New York city taxi drivers. American Economic Review 98 (3), 1069–1082.
Farber, Henry S., Bazerman, Max H., 1986. The general basis of arbitrator behavior: an empirical analysis of conventional and final-offer arbitration. Econometrica 54 (4), 819–844.
Farrell, Joseph, Shapiro, Carl, 1989. Optimal contracts with lock-in. American Economic Review 79 (1), 51–68.
Fehr, E., Goette, L., 2007. Do workers work more when wages are high? Evidence from a randomized field experiment. American Economic Review 97 (1), 298–317.
Fehr, E., List, J., 2004. The hidden costs and returns of incentives—Trust and trustworthiness among CEOs. Journal of the European Economic Association 2 (5), 743–771.
Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition and co-operation. Quarterly Journal of Economics 114, 817–868.
Fehr, Ernst, Schmidt, Klaus M., 2004. Fairness and incentives in a multi-task principal-agent model. Scandinavian Journal of Economics 106 (3), 453–474.

Fehr, E., Fischbacher, U., Tougareva, E., 2002. Do high stakes and competition undermine fairness? Evidence from Russia. University of Zurich, Working Paper No. 120.

Fehr, Ernst, Gächter, Simon, Kirchsteiger, Georg, 1997. Reciprocity as a contract enforcement device: experimental evidence. Econometrica 65 (4), 833–860.

Fehr, E., Klein, A., Schmidt, K., 2007. Fairness and contract design. Econometrica 75 (1), 121–154.

Fehr, Ernst, Hart, Oliver D., Zehnder, Christian, 2008. Contracts as reference points: experimental evidence. Working Paper No. 14501, National Bureau of Economic Research.

Fehr, Ernst, Kirchsteiger, Georg, Riedl, Arno, 1993. Does fairness prevent market clearing? An experimental investigation. Quarterly Journal of Economics 108 (2), 437–459.

Fehr, Ernst, Klein, Alexander, Schmidt, Klaus M., 2007a. Fairness and contract design. Econometrica 75 (1), 121–154.

Fehr, Ernst, Klein, Alexander, Schmidt, Klaus M., 2007b. Adding a stick to the carrot? The interaction of bonuses and fines. American Economic Review Papers and Proceedings 97 (2), 177–181.

Fehr, Ernst, Falk, Armin, 1999. Wage rigidity in a competitive incomplete contract market. Journal of Political Economy 107 (1), 106–134.

Fehr, Ernst, Fischbacher, Urs, von Rosenbladt, Bernhard, Schupp, Jürgen, Wagner, Gert, 2002. A nation-wide laboratory—examining trust and trustworthiness by integrating behavioral experiments into representative surveys. Schmollers Jahrbuch 122, 519–542.

Fehr, Ernst, Gächter, Simon, 2000. Fairness and retaliation: the economics of reciprocity. Journal of Economic Perspectives 14 (3), 159–181.

Fehr, Ernst, Kirchler, Erich, Weichbold, Andreas, Gächter, Simon, 1998. When social norms overpower competition: gift exchange in experimental labor markets. Journal of Labor Economics 16 (2), 324–351.

Fershtman, C., Gneezy, U., 2001. Discrimination in a segmented society: an experimental approach. Quarterly Journal of Economics 116 (1), 351–377.

Fershtman, C., Gneezy, U., Verboven, F., 2005. Discrimination and nepotism: the efficiency of the anonymity rule. Journal of Legal Studies 34 (2), 371–394.

Fershtman, Chaim, Gneezy, Uri, List, John, 2009. Inequity, selfishness and social norms, mimeo.

Ferster, C.B., Skinner, B.F., 1957. Schedules of Reinforcement. Appleton-Century-Crofts, New York.

Frohlich, N., Oppenheimer, J., 1992. Choosing Justice: An Experimental Approach to Ethical Theory. University of California Press, Berkeley, CA.

Forsythe, R., Kennan, J., Sopher, B., 1991a. An experimental analysis of strikes in bargaining games with one-sided private information. American Economic Review 81 (1), 253–278.

Forsythe, Robert, Kennan, John, Sopher, Barry, 1991b. Dividing a shrinking pie: an experimental study of strikes in bargaining games with complete information. In: Isaan, R. Marc (Ed.), Research in Experimental Economics, vol. 4. JAI Press, Greenwich, Conn., London, pp. 223–268.

Frechette, Guillaume, Kagel, John H., Morelli, Massimo, 2005. Behavioral identification in coalitional bargaining: an experimental analysis of demand bargaining and alternating offers. Econometrica 73 (6), 1893–1937.

Frey, B., 1993. Does monitoring increase work effort? The rivalry between trust and loyalty. Economic Inquiry 31, 663–670.

Frey, Bruno S., Oberholzer-Gee, Felix, 1997. The cost of price incentives: an empirical analysis of crowding-out. American Economic Review 87 (4), 746–755.

Fryer, R., Goeree, J., Holt, C., 2005. Experience-based discrimination: classroom games. Journal of Economic Education 36 (2), 160–170.

Gabaix, Xavier, Laibson, David, Moloche, Guillermo, Weinberg, Stephen, 2006. Costly information acquisition: experimental analysis of a boundedly rational model. American Economic Review 96, 1043–1068.

Gächter, Simon, Nosenzo, Daniele, Sefton, Martin, 2008. The impact of social comparisons on reciprocity. CeDEx Discussion Paper No. 2008-09.

Gibbons, R., Waldman, M., 1999. A theory of wage and promotion dynamics inside firms. Quarterly Journal of Economics 114 (4), 1321–1358.

Gill, David, Prowse, Victoria L., 2009. A structural analysis of disappointment aversion in a real effort competition. IZA Working Paper No. 4356.

Georg, Sebastian, Kube, Sebastian, Zultan, Ro'i, 2009. Treating equals unequally: incentives in teams, workers' motivation and production technology. IZA Discussion Paper No. 3959.
Gibbons, Robert, 1987. Piece-rate incentive schemes. Journal of Labor Economics 5 (4), 413–429 (Part 1).
Gibbons, Robert, 1988. Incentives in organizations. Journal of Economic Perspectives 12 (4), 115–132.
Gneezy, U., List, J.A., 2006. Putting behavioral economics to work: testing for gift exchange in labor markets using field experiments. Econometrica 74 (5), 1365–1384.
Gneezy, U., Rustichini, Aldo, 2000. Pay enough or don't pay at all. Quarterly Journal of Economics 115 (3), 791–810.
Gneezy, Uri, Haruvy, Ernan, Roth, Alvin E., 2003. Bargaining under a deadline: evidence from the reverse ultimatum game. In: Special Issue in Honor of Robert W. Rosenthal. Games and Economic Behavior 45 (2), 347–368.
Gneezy, U., Rustichini, A., 2004. Gender and competition at a young age. American Economic Review 94 (2), 377–381.
Gneezy, U., Leonard, K.L., List, J.A., 2009. Gender differences in competition: evidence from a matrilineal and a patriarchal society. Econometrica 77 (5), 1637–1664. http://management.ucsd.edu/faculty/directory/gneezy/docs/gender-differences-competition.pdf.
Gneezy, U., Niederle, M., Rustichini, A., 2003. Performance in competitive environments: gender differences. Quarterly Journal of Economics 118, 1049–1074.
Goodwin, Stephanie A., Williams, Kipling D., Carter-Sowell, Adrienne R., 2010. The psychology of stigma: the costs of attributing ostracism to racism. Journal of Experimental Social Psychology 46, 612–618.
Guth, Werner, Huck, Steffen, Ockenfels, Peter, 1996. Two-level ultimatum bargaining with incomplete information: an experimental study. Economic Journal 106, 593–660.
Guth, Werner, Levati, M. Vittoria, Maciejovsky, Boris, 2005. Deadline effects in sequential bargaining—an experimental study. International Game Theory Review 7 (2), 117–135.
Greenberg, Jerald, 1990. Employee theft as a reaction to underpayment inequity: the hidden cost of pay cuts. Journal of Applied Psychology 75, 561–568.
Grether, David M., Isaac, R. Mark, Plott, Charles R., 1981. The allocation of landing rights by unanimity among competitors. American Economic Review 71 (2), 166–171.
Griffith, Rachel, Neely, Andrew, 2009. Performance pay and managerial experience in multitask teams: evidence from within a firm. Journal of Labor Economics 27 (1), 49–82.
Grout, Paul A., 1984. Investment and wages in the absence of binding contracts: a Nash bargaining approach. Econometrica 52 (2), 449–460.
Gürtler, Oliver, Harbring, Christine, 2007. Feedback in tournaments under commitment problems: theory and experimental evidence. Discussion Paper No. 3111, Institute for the Study of Labor.
Gu, Wulong, Kuhn, Peter, 1998. A theory of holdouts in wage bargaining. American Economic Review 88, 428–449.
Güth, Werner, Schmittberger, Rolf, Schwarze, Bernd, 1982. An experimental analysis of ultimatum bargaining. Journal of Economic Behavior and Organization 3 (4), 367–388.
Güth, W., Königstein, M., Kovács, J., Zala-Mezõ, E., 2001. Fairness within firms: the case of one principal and many agents. Schmalenbach Business Review 53 (2), 82–101.
Hall, Robert E., Lazear, Edward P., 1984. The excess sensitivity of layoffs and quits to demand. Journal of Labor Economics 2 (2), 233–257.
Hamilton, B., Nickerson, J., Owan, H., 2003. Team incentives and worker heterogeneity: an empirical analysis of the impact of teams on productivity and participation. Journal of Political Economy 111 (3), 465–497.
Hamman, John, Loewenstein, George, Weber, Roberto A., 2008. Self-interest through agency: an alternative rationale for the principal-agent relationship. Unpublished Paper, Carnegie Mellon University.
Hannan, R. Lynn, Kagel, John H., Moser, Donald V., 2002. Partial gift exchange in an experimental labor market: impact of subject population differences, productivity differences and effort requests on behavior. Journal of Labor Economics 20 (4), 923–951.
Harbring, Christine, Irlenbusch, Bernd, 2003. An experimental study on tournament design. Labour Economics 10 (4), 443–464.

Harbring, Christine, Irlenbusch, Bernd, 2005. Incentives in tournaments with endogenous prize selection. Journal of Institutional and Theoretical Economics 161 (4), 636–663.
Harbring, Christine, Irlenbusch, Bernd, 2008. How many winners are good to have? On tournaments with sabotage. Journal of Economic Behavior and Organization 65 (3–4), 682–702.
Harbring, C., Irlenbusch, B., Kräkel, M., Selten, R., 2007. Sabotage in corporate contests—an experimental analysis. International Journal of the Economics of Business 14, 201–223.
Harrison, Glenn W., Morgan, Peter, 1990. Search intensity in experiments. Economic Journal 100 (401), 478–486.
Haruvy, Eman, Roth, Alvin E., Utku Unver, M., 2006. The dynamic of law clerk matching: an experimental and computational investigation of proposals for reform of the market. Journal of Economics Dynamics and Control 30 (3), 457–486.
Hashimoto, Masanori, Yu, Ben T., 1980. Specific capital, employment contracts, and wage rigidity. Bell Journal of Economics 11 (2), 536–549.
Hayes, Beth, 1984. Unions and strikes with asymmetric information. Journal of Labor Economics 2 (1), 57–83.
Healy, Paul J., 2007. Group reputations, stereotypes and cooperation in a repeated labor market. American Economic Review 97 (5), 1751–1773.
Hey, John D., 1982. Search for rules for search. Journal of Economic Behavior and Organization 3 (1), 65–81.
Hicks, J.R., 1932. The Theory of Wages. Macmillan, London, Second edition, 1963.
Hoffman, Elizabeth, McCabe, Kevin, Smith, Vernon, 2008. Preferences and property rights in ultimatum and dictator games. In: Plott, Charles R., Smith, Vernon L. (Eds.), Handbook of Experimental Economics Results, vol. 1. North Holland, Amsterdam, pp. 417–422 (Chapter 47).
Holm, Hakan, Engseld, Peter, 2005. Choosing bargaining partners—an experimental study on the impact of information about income, status and gender. Experimental Economics 8 (3), 183–216.
Holm, Håkan, Nystedt, Paul, 2005. Intra-generational trust—a semi-experimental study of trust among different generations. Journal of Economic Behavior and Organization 58 (3), 403–419.
Holmstrom, B., 1982. Moral hazard in teams. Bell Journal of Economics Autumn.
Holmstrom, B., 1999. Managerial incentive problems: a dynamic perspective. Review of Economic Studies 66, 169–182. Originally published in Essays in Honor of Lars Wahlbeck. Helsinki, Finland: 1982.
Holmstrom, B., Milgrom, P., 1991. Multi-task principal-agent problems: incentive contracts, asset ownership, and job design. Journal of Law, Economics and Organization 7, 24–52 (Special issue).
Holt, Charles A., 1995. Industrial organization: a survey of laboratory research. In: Roth, Alvin E., Kagel, John H. (Eds.), The Handbook of Experimental Economics. pp. 349–443.
Holt, C., 2007. Markets, Games, and Strategic Behavior. Pearson Addison Wesley, Boston.
Hvide, H., 2002. Tournament rewards and risk-taking. Journal of Labor Economics 20 (4), 877–898.
Ickes, Barry W., Samuelson, Larry, 1987. Job transfers and incentives in complex organizations: thwarting the ratchet effect. RAND Journal of Economics 18 (2), 275–286.
Irlenbusch, Bernd, Sliwka, Dirk, 2005a. Transparency and reciprocal behavior in employment relations. Journal Economic Behavior and Organization 56 (3), 383–403.
Irlenbusch, Bernd, Sliwka, Dirk, 2005b. Incentives, decision frames and crowding out: an experimental investigation. Discussion Paper No. 1758, Institute for the Study of Labor.
Irlenbusch, Bernd, Sliwka, Dirk, 2006. Career concerns in a simple experimental labour market. European Economic Review 50 (1), 147–170.
Irlenbusch, Bernd, Ruchala, Gabriele K., 2008. Relative rewards within team-based compensation. Labour Economics 15 (2), 141–167.
Isaac, R.M., Walker, J.M., 1988. Group size effects in public goods provision: the voluntary contribution mechanism. Quarterly Journal of Economics 103, 179–200.
Johnson, D.F., Tullar, W., 1972. Style of third party intervention, face-saving, and bargaining behavior. Journal of Experimental Social Psychology 8, 319–330.
Jones, Stephen R.G., 1992. Was there a Hawthorne effect? American Journal of Sociology 98 (3), 451–468.
Kagel, John H., 1987. Economics according to the rats (and pigeons too): what have we learned and what can we hope to learn? In: Roth, Alvin E. (Ed.), Laboratory Experimentation in Economics: Six Points of View. Cambridge University Press, Cambridge.

Kagel, J.H., Battalio, R.C., Green, L., 1995. Economic Choice Theory: An experimental Analysis of Animal Behavior. Cambridge University Press, Cambridge.
Kagel, John H., Roth, A.E., 2000. The dynamics of reorganization in matching markets: a laboratory experiment motivated by a natural experiment. Quarterly Journal of Economics 115 (1), 201–235.
Kahan, J.P., Rapoport, A., Jones, L.V., 1967. Decision making in a sequential search task. Perception and Psychophysics 2, 374–376.
Kahneman, Daniel, Knetsch, Jack, Thaler, Richard, 1986. Fairness and the assumptions of economics. Journal of Business 59, S285–S300.
Kahn, Charles, Huberman, Gur, 1988. Two-sided uncertainty and "up-or-out" contracts. Journal of Labor Economics 6 (4), 423–444.
Kanemoto, Yoshitsugu, MacLeod, W. Bentley, 1992. The ratchet effect and the market for secondhand workers. Journal of Labor Economics 10 (1), 85–98.
Karlan, Dean, 2005. Using experimental economics to measure social capital and predict financial decisions. American Economic Review 95 (5), 1688–1699.
Kelsey, J.E., Allison, J., 1976. Fixed ratio level pressing by UVH rats: work vs. accessibility of sucrose reward. Physiology and Behavior 17 (5), 749–754.
Kennan, John, 1987. The economics of strikes. In: Layard, Richard, Ashenfelter, Orley (Eds.), Handbook of Labor Economics, vol. 2. North-Holland, Amsterdam.
Kennan, John, Wilson, Robert, 1993. Bargaining with private information. Journal of Economic Literature 31 (1), 45–104.
Knez, Marc, Simester, Duncan, 2001. Firm-wide incentives and mutual monitoring at continental airlines. Journal of Labor Economics 19 (4), 743–772.
Koch, Alexander, Morgenstern, Albrecht, Raab, Philippe, 2009. Career concerns incentives: an experimental test. Working Paper No. 2009-1, University of Aarhaus.
Koszegi, B., Rabin, M., 2006. A model of reference-dependent preferences. Quarterly Journal of Economics 121 (4), 1133–1165.
Kritikos, Alexander S., 2006. The impact of compulsory arbitration on bargaining behavior: an experimental study. Economics of Governance 7 (3), 293–315.
Krueger, Alan, Mas, Alexandre, 2004. Strikes, scabs and tread separations: labor strife and the production of defective bridgestone/firestone tires. Journal of Political Economy 112 (2), 253–289.
Kube, Sebastian, Puppe, Clemens, Maréchal, Michel, 2006a. The currency of reciprocity. Unpublished manuscript, University of St. Gallen.
Kube, Sebastian, Puppe, Clemens, Maréchal, Michel, 2006b. Putting reciprocity to work-positive versus negative responses in the field. Unpublished manuscript, University of St. Gallen.
Kübler, Dorothea, Müller, Wieland, Normann, Hans-Theo, 2008. Job market signalling and screening: an experimental comparison. Games and Economic Behavior 64 (1), 219–236.
Kuhn, Michael A., 2009. To settle or not to settle: a review of the literature on arbitration in the laboratory. Unpublished manuscript, University of California, San Diego.
Kuhn, Peter, 1998. Unions and the economy: what we know; what we should know. Canadian Journal of Economics 31 (5), 1033–1056.
Kuhn, Peter, Gu, Wulong, 1999. Learning in sequential wage negotiations: theory and evidence. Journal of Labor Economics 17, 109–140.
Lazear, Edward P., 1989. Pay equality and industrial politics. Journal of Political Economy 97 (3), 561–580.
Lazear, E.P., 2000. Performance pay and productivity. American Economic Review 90 (3), 1346–1361.
Lazear, E., Rosen, S., 1981. Rank-order tournaments as optimum labor contracts. Journal of Political Economy.
Lazear, Edward P., Malmendier, Ulrike, Weber, Roberto A., 2006. Sorting in experiments with application to social preferences. Working Paper No. W12041, National Bureau of Economic Research.
Ledyard, J.O., 1995. Public goods: a survey of experimental research. In: Kagel, J., Roth, A. (Eds.), Handbook of Experimental Economics. Princeton University Press, Princeton.
Ledyard, John, Porter, David, Wessen, Randii, 2000. A market-based mechanism for allocating space shuttle secondary payload priority. Experimental Economics 2 (3), 173–195.
Levi, Ariel, Fried, Yitzhak, 2008. Differences between African Americans and Whites in reactions to affirmative action programs in hiring, promotion, training, and layoffs. Journal of Applied Psychology 93 (5), 1118–1129.

Levitt, S., List, J., 2007. What do laboratory experiments measuring social preferences reveal about the real world? Journal of Economic Perspectives 21 (2), 153–174.
Levitt, S., List, J., 2009. Field experiments in economics: the past, the present, and the future. European Economic Review 53 (1), 1–18.
Levitt, Steven D., List, John A., Reiley, Jr. David H., 2009. What happens in the field stays in the field: exploring whether professionals play minimax in laboratory experiments. NBER Working Paper No. 15609.
Liberman, V., Samuels, S., Ross, L., 2004. The name of the game: predictive power of reputations versus situational labels in determining prisoner's dilemma game moves. Personality and Social Psychology Bulletin 30 (9), 1175–1185.
Lippman, S., McCall, J., 1976a. The economics of job search: a survey—Part I. Economic Inquiry 14, 155–190.
Lippman, S., McCall, J., 1976b. The economics of job search: a survey—Part II. Economic Inquiry 14, 347–389.
List, J., 2008. Homo experimentalis evolves. Science 321 (5886), 207–208.
List, J., 2009. Introduction to field experiments in economics. Journal of Economic Behavior and Organization 70 (3), 439–442.
List, J., Berrens, R., Bohara, A., Kerkvliet, J., 2004. Examining the role of social isolation on stated preferences. American Economic Review 94 (3), 741–752.
Loewenstein, George, Bazerman, Max, Thompson, Leigh, 1989. Social utility and decision making in interpersonal contexts. Journal of Personality and Social Psychology 57, 426–441.
MacLeod, W. Bentley, Malcomson, James M., 1993. Investments, holdup, and the form of market contracts. American Economic Review 83 (4), 811–837.
Malcomson, James M., 1997. Contracts, hold-up, and labor markets. Journal of Economic Literature 35, 1917–1957.
Mas, Alexandre, 2006. Pay, reference points, and police performance. Quarterly Journal of Economics 121 (3), 783–821.
Mas, Alexandre, 2008. Labour unrest and the quality of production: evidence from the construction equipment resale market. Review of Economic Studies 75 (1), 229–258.
Mas, A., Moretti, E., 2009. Peers at work. American Economic Review 99 (1), 112–143.
Mathewson, Stanley B., 1931. Restriction of Output among Unorganized Workers. Viking Press, New York.
Maximiano, Sandra, Sloof, Randolph, Sonnemans, Joep, 2006. Gift exchange and the separation of ownership and control. Discussion Paper No. 2006-037/1, Tinbergen Institute.
Maximiano, Sandra, Sloof, Randolph, Sonnemans, Joep, 2007. Gift exchange in a multi-worker firm. Economic Journal 117 (522), 1025–1050.
McKinney, C. Nicholas, Niederle, Muriel, Roth, Alvin E., 2005. The collapse of a medical labor clearinghouse (and why such failures are rare). American Economic Review 95 (3), 878–889.
McMillan, John, 1994. Selling spectrum rights. Journal of Economic Perspectives 8, 145–162.
Meidinger, Rullière, Jean-Louis, Villeval, Marie-Claire, 2003. Does team-based compensation give rise to problems when agents vary in their ability? Experimental Economics 6 (3), 253–272.
Meyer, Bruce D., 1995. Lessons from the US unemployment insurance experiments. Journal of Economic Literature 33 (1), 91–131.
Meyer, Bruce D., 1996. What have we learned from the Illinois reemployment bonus experiment? Journal of Labor Economics 14 (1), 26–51.
Miller, R., Plott, C., 1985. Product quality signalling in experimental markets. Econometrica 53, 837–872.
Mobius, Markus M., Rosenblat, Tanya S., 2006. Why beauty matters. American Economic Review 96 (1), 222–235.
Moffitt, Robert, Nicholson, Walter, 1982. The effect of unemployment insurance on unemployment: the case of federal supplemental benefits. Review of Economics and Statistics 64 (1), 1–11.
Mongell, Susan, Roth, Alvin E., 1991. Sorority rush as a two-sided matching mechanism. American Economic Review 81 (3), 441–464.
Mortensen, Dale, Pissaridies, Christopher, 1994. Job creation and job destruction in the theory of unemployment. Review of Economic Studies 61 (3), 397–415.

Munier, Bertrand, Costin, Zaharia, 2002. High stakes and acceptance behavior in ultimatum bargaining: a contribution from an international experiment. Theory and Decision 53 (3), 187–207.
Murninghan, J.K., 2008a. A general model for experimental inquiry in economics and social psychology, mimeo.
Murnighan, J. Keith, 2008b. Fairness in ultimatum bargaining. In: Plott, Charles R., Smith, Vernon L. (Eds.), Handbook of Experimental Economics Results, vol. 1. North Holland, Amsterdam, pp. 436–453 (Chapter 50).
Nalbantian, Haig R., Schotter, Andrew, 1995. Matching and efficiency in the baseball free-agent system: an experimental examination. Journal of Labor Economics 13 (1), 1–31.
Nalbantian, Haig R., Schotter, Andrew, 1997. Productivity under group incentives: and experimental study. American Economic Review 87 (3), 314–341.
Nash, John, 1950. The bargaining problem. Econometrica 18, 155–162.
Nash, John, 1953. Two-person cooperative games. Econometrica 21, 128–140.
Neale, Margaret A., Bazerman, Max H., 1983. The role of perspective-taking ability in negotiating and under different forms of arbitration. Industrial & Labor Relations Review 36 (3), 378–388.
Niederle, Muriel, Vesterlund, Lise, 2007. Do women shy away from competition? Do men compete too much? Quarterly Journal of Economics 122 (3), 1067–1101.
Niederle, Muriel, Yestrumskas, Alexandra H., 2008. Gender differences in seeking challenges: the role of institutions. Working Paper No. 13922, National Bureau of Economic Research.
Niederle, Muriel, Roth, Alvin E., 2003. Unraveling reduces mobility in a labor market: gastroenterology with and without a centralized match. Journal of Political Economy 111 (6), 1342–1352.
Niederle, Muriel, Roth, Alvin E., 2009. Market culture: how rules governing exploding offers affect market performance. American Economic Journal: Microeconomics 1 (2), 199–219.
Niederle, Muriel, Roth, Alvin E., Unver, M. Utku, 2009. Unraveling results from comparable demand and supply: an experimental investigation. National Bureau of Economic Research, Inc., NBER Working Papers: 15006.
Niederle, Muriel, Segal, Carmit, Vesterlund, Lise, 2008. How costly is diversity? Affirmative action in light of gender differences in competitiveness. Working Paper No. W13923, National Bureau of Economic Research.
Nieken, Petra, Sliwka, Dirk, 2010. Risk-taking tournaments: theory and experimental evidence. Journal of Economic Psychology 31 (3), 254–268.
Notz, William W., Starke, Frederick A., 1978. Final-offer versus conventional arbitration as means of conflict management. Administrative Science Quarterly 23 (2), 189–203.
Ochs, Jack, Roth, Alvin E., 1989. An experimental study of sequential bargaining. American Economic Review 79 (3), 355–384.
Ockenfels, Axel, Selten, Reinhard, 2000. An experiment on the hypothesis of involuntary truth-signalling in bargaining. Games and Economic Behavior 33 (1), 90–116.
Oettinger, Gerald S., 1999. An empirical analysis of the daily labor supply of stadium vendors. Journal of Political Economy 107 (2), 360–392.
Offerman, Theo, 2002. Hurting hurts more than helping helps. European Economic Review 46 (8), 1423–1437.
O'Keefe, M., Viscusi, W.K., Zeckhauser, R.J., 1984. Economic contests: comparative reward schemes. Journal of Labor Economics.
Oosterbeek, Hessel, Sloof, Randolph, Sonnemans, Joep, 2006. Rent-seeking versus productive activities in a multi-task experiment. Discussion Paper No. 06-083/1, Tinbergen Institute.
Oosterbeek, Hessel, Sloof, Randolph, Sonnemans, Joep, 2007a. Who should invest in specific training? Journal of Population Economics 20 (2), 329–357.
Oosterbeek, Hessel, Sloof, Randolph, Sonnemans, Joep, 2007b. Promotion rules and skill acquisition: an experimental study. Economica 74 (294), 259–297.
Osborne, Martin, Rubinstein, Ariel, 1990. Bargaining and Markets. Academic Press, Boston.
Owens, M.F., Kagel, J.H., 2010. Minimum wage restrictions and employee effort in incomplete labor markets: an experimental investigation. Journal of Economic Behavior and Organization 73 (3), 317–326.
Palacios-Huertas, I., Volij, O., 2008. Experientia docet: professionals play minimax in laboratory experiments. Econometrica 76 (1), 71–115.

Plott, Charles R., 1987. Dimensions of parallelism: some policy applications of experimental methods. In: Roth, Alvin E. (Ed.), Laboratory Experimentation in Economics: Six Points of View. Cambridge University Press, Cambridge.
Plott, Charles R., Smith, Vernon L. (Eds.), 2008. Handbook of Experimental Economics Results, vol. 1. North Holland, Amsterdam, pp. 417–422.
Prendergast, Canice, 1993. The role of promotion in inducing specific human capital acquisition. Quarterly Journal of Economics 108 (2), 523–534.
Rabin, Matthew, 1993. Incorporating fairness into game theory and economics. American Economic Review 83, 1281–1302.
Rapaport, A., 1995. Individual strategies in a market-entry game. Group Decision and Negotiation 4, 117–133.
Rapoport, A., Tversky, A., 1966. Cost and accessibility of offers as determinants of optimal stopping. Psychonomic Science 4, 145–146.
Rapoport, A., Tversky, A., 1970. Choice behavior in an optimal stopping task. Organizational Behavior and Human Performance 5, 105–120.
Rassenti, S.J, Smith, V.L, Bulfin, R.L., 1982. A combinational auction mechanism for airport time slot allocation. Bell Journal of Economics 13 (2), 402–417.
Roth, Alvin E., 1984. The evolution of the labor market for medical interns and residents: a case study in game theory. Journal of Political Economy 92 (6), 991–1016.
Roth, Alvin E., 1986. On the allocation of residents to rural hospitals: a general property of two-sided matching markets. Econometrica 54 (2), 425–427.
Roth, Alvin E., 1995. Bargaining experiments. In: Kagel, John H., Roth, Alvin E. (Eds.), The Handbook of Experimental Economics, vol. 1. Princeton University Press, pp. 253–348 (Chapter 5).
Roth, Alvin E., 2002. The economist as engineer: game theory, experimentation, and computation as tools for design economics. Econometrica 70 (4), 1341–1378.
Roth, Alvin E., 2008. What have we learned from market design? Economic Journal 118 (527), 285–310.
Roth, Alvin E., Peranson, Elliott, 1999. The redesign of the matching market for American physicians: some engineering aspects of economic design. American Economic Review 89 (4), 748–780.
Roth, Alvin E., Xing, Xiaolin, 1994. Jumping the gun: imperfections and institutions related to the timing of market transactions. American Economic Review 84 (4), 992–1044.
Roth, Alvin E., Prasnikar, Vesna, Okuno-Fujiwara, Masahiro, Zamir, Shmuel, 1991. Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: an experimental study. American Economic Review 81 (5), 1068–1095.
Roth, Alvin E., Sonmez, Tayfun, Unver, M. Utku, 2007. Efficient kidney exchange: coincidence of wants in markets with compatibility-based preferences. American Economic Review 97 (3), 828–851.
Roth, Alvin E., Sotomayor, Marilda, 1989. The college admissions problem revisited. Econometrica 57 (3), 559–570.
Roth, Alvin E., Xing, Xiaolin, 1997. Turnaround time and bottlenecks in market clearing: decentralized matching in the market for clinical psychologists. Journal of Political Economy 105 (2), 284–329.
Rubinstein, Ariel, 1982. Perfect equilibrium in a bargaining model. Econometrica 50, 97–109.
Runkel, P., McGrath, J., 1972. Research on Human Behavior: A Systematic Guide. Holt, Rinehart & Winston, New York.
Rütstrom, E., Williams, M., 2000. Entitlements and fairness: an experimental study of distributive preferences. Journal of Economic Behavior and Organization 43, 75–89.
Selten, R., 1967. Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments. In: Sauermann, H. (Ed.), Beiträge zur experimentellen Wirtschaftsforschung. Mohr, Tübingen, pp. 136–168.
Shiller, R.J., 1993. Macro Markets: Creating Institutions for Managing Society's Largest Economic Risks. In: Clarendon Lectures in Economics, Oxford University Press, Oxford.
Schminki, Marshall, Cropanzano, Russell, Rupp, Deborah, 2002. Organizational structure and fairness perceptions: the moderating effects of organizational level. Organizational Behavior and Human Decision Processes 89, 881–905.
Schnedler, Wendelin, Vadovic, Radovan, 2007. Legitimacy of control. Discussion Paper No. 3013, Institute for the Study of Labor.

Schotter, Andrew, Braunstein, Yale M., 1981. Economic search: an experimental study. Economic Inquiry 19 (1), 1–25.
Schotter, Andrew, Braunstein, Yale M., 1982. Labor market search: an experimental study. Economic Inquiry 20 (1), 133–144.
Schotter, Andrew, Weigelt, Keith, 1992. Asymmetric tournaments, equal opportunity laws and affirmative action: some experimental results. Quarterly Journal of Economics 107 (2), 511–539.
Schotter, Andrew, Zheng, Wei, Snyder, Blaine, 2000. Bargaining through agents: an experimental study of delegation and commitment. Games and Economic Behavior 30 (2), 248–292.
Schram, Arthur, Brandts, Jordi, Gërxhani, Klarita, 2007. Information networks and worker recruitment. Unpublished Paper, University of Amsterdam.
Schunk, Daniel, 2009. Behavioral heterogeneity in dynamic search situations: theory and experimental evidence. Journal of Economic Dynamics and Control 33 (9), 1719–1738.
Schunk, Daniel, Winter, Joachim, 2009. The relationship between risk attitudes and heuristics in search tasks: a laboratory experiment. Journal of Economic Behavior and Organization 71 (2), 347–360.
Schwieren, C., 2003. The gender-wage gap – due to differences in efficiency wage effects or discrimination? Research Memoranda 046, Maastricht: METEOR, Maastricht Research School of Economics of Technology and Organization.
Shapiro, Carl, Stiglitz, J.E., 1984. Equilibrium unemployment as a worker discipline device. American Economic Review 74, 433–444.
Shearer, Bruce, 2004. Piece rates, fixed wages and incentives: evidence from a field experiment. Review of Economic Studies 71 (2), 513–534.
Siegel, Sidney, Fouraker, Lawrence E., 1960. Bargaining and Group Decision Making: Experiments in Bilateral Monopoly. McGraw-Hill, New York.
Sillamaa, M.A., 1999a. How work effort responds to wage taxation: a non-linear versus a linear tax experiment. Journal of Economic Behavior and Organization 39 (2), 219–233.
Sillamaa, M.A., 1999b. How work effort responds to wage taxation: an experimental test of a zero top marginal tax rate. Journal of Public Economics 73 (1), 125–134.
Skarlicki, Daniel P., Folger, Robert, 1997. Retaliation in the workplace: the roles of distributive, procedural, and interactional justice. Journal of Applied Psychology 82, 434–443.
Slade, Margaret E., 1996. Multitask agency and contract choice: an empirical exploration. International Economic Review 37 (2), 465–486.
Slonim, R., Roth, A.E., 1998. Learning in high stakes ultimatum games: an experiment in the Slovak Republic. Econometrica 66 (3), 569–596.
Sloof, Randolph, Oosterbeek, Hessel, Sonnemans, Joep, 2007. Does making specific investments unobservable boost investment incentives? Journal of Economics and Management Strategy 16 (4), 911–942.
Sonnemans, Joep, 1998. Strategies of search. Journal of Economic Behavior and Organization 35 (3), 309–332.
Sonnemans, Joep, 2000. Decisions and strategies in a sequential search experiment. Journal of Economic Psychology 21 (1), 91–102.
Sopher, Barry, 1990. Bargaining and the joint-cost theory of strikes: an experimental study. Journal of Labor Economics 8 (1), 48–74 (Part 1).
Sprinkle, Geoffrey B., 2000. The effect of incentive contract on learning and performance. Accounting Review 75 (3), 299–326.
Starke, Frederick A., Notz, William W., 1981. Pre- and post-intervention effects of conventional versus final offer arbitration. Academy of Management Journal 24 (4), 832–850.
Sterbenz, Frederic P., Phillips, Owen, R., 2001. Bargaining experiments with deadlines and random delays. Economic Inquiry 39 (4), 616–626.
Sutter, Matthias, Kocher, Martin, 2007. Trust and trustworthiness across different age groups. Games and Economic Behavior 59, 364–382.
Sutter, Matthias, Strassmair, Christina, 2009. Communication, cooperation and collusion in team tournaments: an experimental study. Games and Economic Behavior 66 (1), 506–525.
Svejnar, J., Singer, M., 1994. Using vouchers to privatize an economy: the Czech and Slovak case. Economics of Transition 2, 43–69.

Swenson, C.W., 1988. Taxpayer behavior in response to taxation: an experimental analysis. Journal of Accounting and Public Policy 7, 1–28.
Tournadre, Fabienne, Villeval, Marie-Claire, 2004. Learning from strikes. Labour Economics 11 (2), 243–264.
Van Dijk, Frans, Sonnemans, Joep, Van Winden, Frans, 2001. Incentive systems in a real effort experiment. European Economic Review 45 (2), 187–214.
Van Huyck, John B., Battalio, Raymond C., Beil, Richard O., 1990. Tacit coordination games, strategic uncertainty, and coordination failure. American Economic Review 80 (1), 234–248.
Vandegrift, Donald, Brown, Paul, 2003. Task difficulty, incentive effects and the selection of high-variance strategies: an experimental examination of tournament behavior. Labour Economics 10 (4), 481–497.
Vandegrift, Donald, Yavas, Abdullah, Brown, Paul M., 2007. Incentive effects and overcrowding in tournaments: an experimental analysis. Experimental Economics 10 (4), 345–368.
Von Siemens, F., 2004. Social preferences, sorting, and employment contracts. Chapter 2 of Ph.D. Dissertation, University of Munich.
Weber, Roberto A., 2006. Managing growth to achieve efficient coordination in large groups. American Economic Review 96 (1), 114–126.
Williamson, Oliver, 1985. The Economic Institutions of Capitalism. The Free Press, New York.
Wilson, Rick K., Eckel, Catherine C., 2006. Judging a book by its cover: beauty and expectations in the trust game. Political Research Quarterly 59 (2), 189–202.
Wilson, R.B., 2002. Architecture of power markets. Econometrica 70, 1299–1340.
Winter, Eyal, 2004. Incentives and discrimination. American Economic Review 94 (3), 764–773.

CHAPTER 4

The Structural Estimation of Behavioral Models: Discrete Choice Dynamic Programming Methods and Applications

Michael P. Keane [*] **, Petra E. Todd** [**]**, Kenneth I. Wolpin** [**]

[*] University of Technology, Sydney and Arizona State University

[**] University of Pennsylvania

Contents

Abstract

The purpose of this chapter is twofold: (1) to provide an accessible introduction to the methods of structural estimation of discrete choice dynamic programming (DCDP) models and (2) to survey the contributions of applications of these methods to substantive and policy issues in labor economics.

Handbook of Labor Economics, Volume 4a
ISSN 0169-7218, DOI 10.1016/S0169-7218(11)00410-2

The first part of the chapter describes solution and estimation methods for DCDP models using, for expository purposes, a prototypical female labor force participation model. The next part reviews the contribution of the DCDP approach to three leading areas in labor economics: labor supply, job search and human capital. The final section discusses approaches to validating DCDP models.

JEL classification: J; C51; C52; C54

Keywords: Structural estimation; Discrete choice; Dynamic programming; Labor supply; Job search; Human capital

1. INTRODUCTION

The purpose of this chapter is twofold: (1) to provide an accessible introduction to the methods of structural estimation of discrete choice dynamic programming (DCDP) models and (2) to survey the contributions of applications of these methods to substantive and policy issues in labor economics.[1] The development of estimation methods for DCDP models over the last 25 years has opened up new frontiers for empirical research in labor economics as well as other areas such as industrial organization, economic demography, health economics, development economics and political economy.[2] Reflecting the generality of the methodology, the first DCDP papers, associated with independent contributions by Gotz and McCall (1984), Miller (1984), Pakes (1986), Rust (1987) and Wolpin (1984), addressed a variety of topics, foreshadowing the diverse applications to come in labor economics and other fields. Gotz and McCall considered the sequential decision to re-enlist in the military, Miller the decision to change occupations, Pakes the decision to renew a patent, Rust the decision to replace a bus engine and Wolpin the decision to have a child.

The first part of this chapter provides an introduction to the solution and estimation methods for DCDP models. We begin by placing the method within the general latent variable framework of discrete choice analysis. This general framework nests static and dynamic models and nonstructural and structural estimation approaches. Our discussion of DCDP models starts by considering an agent making a binary choice. For concreteness, and for simplicity, we take as a working example the unitary model of a married couple's decision about the woman's labor force participation. To fix ideas, we use the static model with partial wage observability, that is, when wage offers are observed only for women who are employed, to draw the connection between theory, data and estimation approaches. In that context, we delineate several goals of estimation, for example, testing theory or evaluating counterfactuals, and discuss the ability of alternative estimation approaches, encompassing those that are parametric or nonparametric and

[1] More technical discussions can be found in the surveys by Rust (1993, 1994), Miller (1997) and Aguirregebaria and Mira (forthcoming), as well as in a number of papers cited throughout this chapter.

[2] Their use has spread to areas outside of traditional economics, such as marketing, in which it is arguably now the predominant approach to empirical research.

structural or nonstructural, to achieve those goals. We show how identification issues relate to what one can learn from estimation.

The discussion of the static model sets the stage for dynamics, which we introduce again, for expository purposes, within the labor force participation example by incorporating a wage return to work experience (learning by doing).[3] A comparison of the empirical structure of the static and dynamic models reveals that the dynamic model is, in an important sense, a static model in disguise. In particular, the essential element in the estimation of both the static and dynamic model is the calculation of a latent variable representing the difference in payoffs associated with the two alternatives (in the binary case) that may be chosen. In the static model, the latent variable is the difference in alternative-specific utilities. In the case of the dynamic model, the latent variable is the difference in alternative-specific value functions (expected discounted values of payoffs). The only essential difference between the static and dynamic cases is that alternative-specific utilities are more easily calculated than alternative-specific value functions, which require solving a dynamic programming problem. In both cases, computational considerations play a role in the choice of functional forms and distributional assumptions.

There are a number of modeling choices in all discrete choice analyses, although some are more important in the dynamic context because of computational issues. Modeling choices include the number of alternatives, the size of the state space, the error structure and distributional assumptions and the functional forms for the structural relationships. In addition, in the dynamic case, one must make an assumption about how expectations are formed.[4] To illustrate the DCDP methodology, the labor force participation model assumes additive, normally distributed, iid over time errors for preferences and wage offers. We first discuss the role of exclusion restrictions in identification, and work through the solution and estimation procedure. We then show how a computational simplification can be achieved by assuming errors to be independent type 1 extreme value (Rust, 1987) and describe the model assumptions that are consistent with adopting that simplification. Although temporal independence of the unobservables is often assumed, the DCDP methodology does not require it. We show how the solution and estimation of DCDP models is modified to allow for permanent unobserved heterogeneity and for serially correlated errors. In the illustrative model, the state space was chosen to be of a small finite dimension. We then describe the practical problem that arises in implementing the DCDP methodology as the state space expands, the well-known curse of dimensionality (Bellman, 1957), and describe suggested practical solutions found in the literature including discretization, approximation and randomization.

[3] Most applications of DCDP models assume that agents, usually individuals or households, solve a finite horizon problem in discrete time. For the most part, we concentrate on that case and defer discussion of infinite horizon models to the discussion of the special case of job search models. We do not discuss continuous time models except in passing.

[4] The conventional approach assumes that agents have rational expectations. An alternative approach directly elicits subjective expectations (see, e.g., Dominitz and Manski, 1996, 1997; Van der Klaauw, 2000; Manski, 2004).

To illustrate the DCDP framework in a multinomial choice setting, we extend the labor force participation model to allow for a fertility decision at each period and for several levels of work intensity. In that context, we also consider the implications of introducing nonadditive errors (that arise naturally within the structure of models that fully specify payoffs and constraints) and general functional forms. It is a truism that any dynamic optimization model that can be (numerically) solved can be estimated.

Throughout the presentation, the estimation approach is assumed to be maximum likelihood or, as is often the case when there are many alternatives, simulated maximum likelihood. However, with simulated data from the solution to the dynamic programming problem, other methods, such as minimum distance estimation, are also available. We do not discuss those methods because, except for solving the dynamic programming model, their application is standard. Among the more recent developments in the DCDP literature is a Bayesian approach to the solution and estimation of DCDP models. Although the method has the potential to reduce the computational burden associated with DCDP models, it has not yet found wide application. We briefly outline the approach. All of these estimation methods require that the dynamic programming problem be fully solved (numerically). We complete the methodology section with a brief discussion of a method that does not require solving the full dynamic programming problem (Hotz and Miller, 1993).

Applications of the DCDP approach within labor economics have spanned most major areas of research. We discuss the contributions of DCDP applications in three main areas: (i) labor supply, (ii) job search and (iii) schooling and career choices. Although the boundaries among these areas are not always clear and these areas do not exhaust all of the applications of the method in labor economics, they form a reasonably coherent taxonomy within which to demonstrate key empirical contributions of the approach.[5] In each area, we show how the DCDP applications build on the theoretical insights and empirical findings in the prior literature. We highlight the findings of the DCDP literature, particularly those that involve counterfactual scenarios or policy experiments.

The ambitiousness of the research agenda that the DCDP approach can accommodate is a major strength. This strength is purchased at a cost. To be able to perform counterfactual analyses, DCDP models must rely on extra-theoretic modeling choices, including functional form and distributional assumptions. Although the DCDP approach falls short of an assumption-free ideal, as do all other empirical approaches, it is useful to ask whether there exists convincing evidence about the credibility of these exercises. In reviewing the DCDP applications, we pay careful attention to the model validation exercises that were performed. The final section of the chapter addresses the overall issue of model credibility.

[5] A notable omission is the literature on retirement behavior. Although that literature relies heavily on the DCDP approach, the previous Handbook of Labor Economics chapter by Lumsdaine and Mitchell provides an extensive survey up to that time. We decided to concentrate on DCDP literature that to date has not been surveyed in the Handbook.

2. THE LATENT VARIABLE FRAMEWORK FOR DISCRETE CHOICE PROBLEMS

The development of the DCDP empirical framework was a straightforward and natural extension of the static discrete choice framework. The common structure they share is based on the latent variable specification, the building block for all economic models of discrete choice. To illustrate the general features of the latent variable specification, consider a binary choice model in which an economic agent with imperfect foresight, denoted by i, makes a choice at each discrete period t, from $t = 1, \ldots, T$, between two alternatives $d_{it} \in \{0, 1\}$. In the labor economics context, examples might be the choice of whether to accept a job offer or remain unemployed or whether to attend college or enter the labor force. The outcome is determined by whether a latent variable, v_{it}^*, reflecting the difference in the (expected) payoffs of the $d_{it} = 1$ and $d_{it} = 0$ alternatives, crosses a scalar threshold value, which, without loss of generality, is taken to be zero. The preferred alternative is the one with the largest payoff, i.e., where $d_{it} = 1$ if $v_{it}^* \geq 0$ and $d_{it} = 0$ otherwise.

In its most general form, the latent variable may be a function of three types of variables: $\widetilde{D}_{it}$, a vector of the history of past choices $(d_{i\tau} : \tau = 1, \ldots, t-1)$, $\widetilde{X}_{it}$, a vector of contemporaneous and lagged values of J additional variables $(X_{ij\tau} : j = 1, \ldots, J; \tau = 1, \ldots, t)$ that enter the decision problem, and $\widetilde{\epsilon}_{it}$ $(\epsilon_{i\tau} : \tau = 1, \ldots, t)$, a vector of contemporaneous and lagged unobservables that also enter the decision problem.[6] The agent's decision rule at each age is given by whether the latent variable crosses the threshold, that is,

$$\begin{aligned} d_{it} = 1 \quad & \text{if } v_{it}^*(\widetilde{D}_{it}, \widetilde{X}_{it}, \widetilde{\epsilon}_{it}) \geq 0, \\ = 0 \quad & \text{if } v_{it}^*(\widetilde{D}_{it}, \widetilde{X}_{it}, \widetilde{\epsilon}_{it}) < 0. \end{aligned} \tag{1}$$

All empirical binary choice models, dynamic or static, are special cases of this formulation. The underlying behavioral model that generated the latent variable is dynamic if agents are forward looking *and* either v_{it}^* contains past choices, $\widetilde{D}_{it}$, or unobservables, $\widetilde{\epsilon}_{it}$, that are serially correlated.[7] The underlying model is static (i) if agents are myopic or (ii) if agents are forward looking and there is no link among the past, current and future periods through $\widetilde{D}_{it}$ or serially correlated unobservables.

Researchers may have a number of different, though not necessarily mutually exclusive, goals. They include:

[6] As will be seen in the empirical applications we consider, there are a wide range of types of variables that would be included in X. Their common feature is that they are not directly choices of the agent, although they may be affected by prior choices or correlated with choices without being directly affected by them.

[7] By forward looking, we simply mean that agents take into account the effect of their current actions on future welfare. How exactly they form expectations about the impact of those actions and about future preferences and constraints are specific modeling choices.

1. Test a prediction of the theory, that is, how an observable variable in v_{it}^* affects d_{it}.
2. Determine the affect of a change in $\widetilde{D}_{it}$ or $\widetilde{X}_{it}$ on choices (either within or outside of the sample variation).
3. Determine the affect of a change in something not in $\widetilde{D}_{it}$ or $\widetilde{X}_{it}$ on choices, that is, in something that does not vary in the sample.

It is assumed that these statements are ceteris paribus, not only in the sense of conditioning on the other observables, but also in conditioning on the unobservables and their joint conditional (on observables) distribution.[8] Different empirical strategies, for example, structural or nonstructural, may be better suited for some of these goals than for others.

3. THE COMMON EMPIRICAL STRUCTURE OF STATIC AND DYNAMIC DISCRETE CHOICE MODELS

In drawing out the connection between the structure of static and dynamic discrete choice models, it is instructive to consider an explicit example. We take as the prime motivating example one of the oldest and most studied topics in labor economics, the labor force participation of married women.[9] We first illustrate the connection between research goals and empirical strategies in a static framework and then modify the model to allow for dynamics.

3.1. Married woman's labor force participation

3.1.1. Static model

Consider the following static model of the labor force participation decision of a married woman. Assume a unitary model in which the couple's utility is given by

$$U_{it} = U(c_{it}, 1 - d_{it}; n_{it}(1 - d_{it}), \kappa_{it}(1 - d_{it}), \epsilon_{it}(1 - d_{it})), \tag{2}$$

where c_{it} is household i's consumption at period t, $d_{it} = 1$ if the wife works and is equal to zero otherwise, n_{it} is the number of young children in the household, and κ_{it} are other observable factors and ϵ_{it} unobservable factors that affect the couple's valuation of the wife's leisure (or home production). In this context, t corresponds to the couple's duration of marriage. The utility function has the usual properties: $\partial U/\partial C > 0, \partial^2 U/\partial C^2 < 0, U(C, 1) > U(C, 0)$.

The wife receives a wage offer of w_{it} in each period t and the husband, who is assumed to work each period, generates income y_{it}. If the wife works, the household

[8] By maintaining the same joint distribution when performing the ceteris paribus change, we are assuming that the change in an observable variable does not induce a change in the joint distribution of unobservables. This assumption is not the same as assuming conditional independence.

[9] The modern approach to this topic began with Mincer (1962).

incurs a per-child child-care cost, π, which is assumed to be time-invariant and the same for all households.[10] The household budget constraint is thus

$$c_{it} = y_{it} + w_{it}d_{it} - \pi n_{it}d_{it}. \tag{3}$$

Wage offers are not generally observed for nonworkers. It is, thus, necessary to specify a wage offer function to carry out estimation. Let wage offers be generated by

$$w_{it} = w(z_{it}, \eta_{it}), \tag{4}$$

where z_{it} are observable and η_{it} unobservable factors. z_{it} would conventionally contain educational attainment and "potential" work experience (age − education − 6). Unobservable factors that enter the couple's utility function (ϵ_{it}) and unobservable factors that influence the woman's wage offer (η_{it}) are assumed to be mutually serially uncorrelated and to have joint distribution $F_{\epsilon,\eta|y,\kappa,z,n}$.

Substituting (3) into (2) using (4) yields

$$\begin{aligned} U_{it} = U(&y_{it} + w(z_{it}, \eta_{it})d_{it} \\ &- \pi n_{it}d_{it}, 1 - d_{it}; n_{it}(1 - d_{it}), \kappa_{it}(1 - d_{it}), \epsilon_{it}(1 - d_{it})), \end{aligned} \tag{5}$$

from which we get alternative-specific utilities, U_{it}^1 if the wife works and U_{it}^0 if she does not, namely

$$U_{it}^1 = U(y_{it} + w(z_{it}, \eta_{it}) - \pi n_{it}, 0), \qquad U_{it}^0 = U(y_{it}, 1; n_{it}, \kappa_{it}, \epsilon_{it}). \tag{6}$$

The latent variable function, the difference in utilities, $U_{it}^1 - U_{it}^0$, is thus given by

$$v_{it}^* = v^*(y_{it,}z_{it}, n_{it}, \kappa_{it}, \epsilon_{it}, \eta_{it}). \tag{7}$$

The participation decision is determined by the sign of the latent variable: $d_{it} = 1$ if $v_{it}^* \geq 0$, $d_{it} = 0$ otherwise.

It is useful to distinguish the household's state space, Ω_{it}, consisting of all of the determinants of the household's decision, that is, $y_{it,}z_{it}, n_{it}, \kappa_{it}, \epsilon_{it}, \eta_{it}$, from the part of the state space observable to the researcher, Ω_{it}^-, that is, consisting only of $y_{it,}z_{it}, n_{it}, \kappa_{it}$. Now, define $S(\Omega_{it}^-) = \{\epsilon_{it}, \eta_{it}|v^*(\epsilon_{it}, \eta_{it}; \Omega_{it}^-) > 0\}$ to be the set of values of the unobservables that enter the utility and wage functions that induces a couple with a given observable state space (Ω_{it}^-) to choose $d_{it} = 1$. Then, the probability of choosing $d_{it} = 1$, conditional on Ω_{it}^-, is given by

[10] We treat the price of child care as parametric in part to illustrate how alternative approaches to estimation are related to achieving goal 3. A more complete model would allow for a choice among alternative types of child care, for example, of varying qualities, which differ in their price and which may vary over time.

$$\Pr(d_{it} = 1|\Omega_{it}^{-}) = \int_{S(\Omega_{it}^{-})} \mathrm{d}F_{\epsilon,\eta|y,\kappa,z,n} = G(y_{it}, z_{it}, n_{it}, \kappa_{it}), \tag{8}$$

where $\Pr(d_{it} = 0|\Omega_{it}^{-}) = 1 - \Pr(d_{it} = 1|\Omega_{it}^{-})$.

As is clear from (8), $G(y_{it}, z_{it}, n_{it}, \kappa_{it})$ is a composite of three elements of the model: $U(\cdot)$, $w(\cdot)$, $F_{\epsilon,\eta|y,\kappa,z,n}$. These elements comprise the *structure* of the participation model. Structural estimation (S) is concerned with recovering some or all of the structural elements of the model. Nonstructural (NS) estimation is concerned with recovering $G(\cdot)$. In principal, each of these estimation approaches can adopt auxiliary assumptions in terms of parametric (P) forms for some or all of the structural elements or for $G(\cdot)$ or be nonparametric (NP). Thus, there are four possible approaches to estimation: NP-NS, P-NS, NP-S and P-S.[11]

We now turn to a discussion about the usefulness of each of these approaches for achieving the three research goals mentioned above. The first research goal, testing the theory, requires that there be at least one testable implication of the model. From (6) and the properties of the utility function, it is clear that an increase in the wage offer increases the utility of working, but has no effect on the utility of not working. Thus, the probability of working for any given agent must be increasing in the wage offer. The second goal, to determine the impact of changing any of the state variables in the model on an individual's participation probability, requires taking the derivative of the participation probability with respect to the state variable of interest. The third goal requires taking the derivative of the participation probability with respect to something that does not vary in the data. That role is played by the unknown child care cost parameter, π. Determining its impact would provide a quantitative assessment of the effect of a child care subsidy on a married woman's labor force participation.[12]

Given the structure of the model, to achieve any of these goals, regardless of the estimation approach, it is necessary to adopt an assumption of independence between the unobservable factors affecting preferences and wage offers and the observable factors. Absent such an assumption, variation in the observables, $y_{it}, z_{it}, n_{it}, \kappa_{it}$, either among individuals or over time for a given individual, would cause participation to differ both because of their effect on preferences and/or wage offers and because of their relationship to the unobserved determinants of preferences and/or wage offers through $F_{\epsilon,\eta|y,\kappa,z,n}$. In what follows, we adopt the assumption of full independence, that is, $F_{\epsilon,\eta|y,\kappa,z,n} = F_{\epsilon,\eta}$, so as not to unduly complicate the discussion.

Nonparametric, nonstructural

If we make no further assumptions, we can estimate $G(\cdot)$ nonparametrically.

[11] In this taxonomy, semi-parametric and semi-structural categories fall into the parametric (P) and structural (S) categories.

[12] As before (see footnote 8), we assume that the change in an observable variable does not induce a change in the joint distribution of unobservables.

Goal 1: To accomplish the first goal, we need to be able to vary the wage offer independently of other variables that affect participation. To do that, there must be an exclusion restriction, in particular, a variable in z_{it} that is not in κ_{it}. Moreover, determining the sign of the effect of a wage increase on the participation probability requires knowing the sign of the effect of the variable in z_{it} (not in κ_{it}) on the wage. Of course, if we observed all wage offers, the wage would enter into the latent variable rather than the wage determinants (z_{it} and η_{it}) and the prediction of the theory could be tested directly without an exclusion restriction.

What is the value of such an exercise? Assume that the observation set is large enough that sampling error can be safely ignored and consider the case where all wage offers are observed. Suppose one finds, after nonparametric estimation of the participation probability function, that there is some "small" range of wages over which the probability of participation is declining as the wage increases. Thus, the theory is rejected by the data. Now, suppose we wanted to use the estimated participation probability function to assess the impact of a proportional wage tax on participation. This is easily accomplished by comparing the sample participation probability in the data with the participation probability that comes about by reducing each individual's wage by the tax. Given that the theory is rejected, should we use the participation probability function for this purpose? Should our answer depend on how large is the range of wages over which the violation occurs? Should we add more κ_{it} variables and retest the model? And, if the model is not rejected after adding those variables, should we then feel comfortable in using it for the tax experiment? If there are no ready answers to these questions in so simple a model, as we believe is the case, then how should we approach them in contexts where the model's predictions are not so transparent and therefore for practical purposes untestable, as is normally the case in DCDP models? Are there other ways to validate models? We leave these as open questions for now, but return to them in the concluding section of the chapter.

Goal 2: Clearly, it is possible, given an estimate of G, to determine the effect on participation of a change in any of the variables within the range of the data. However, one cannot predict the effect of a change in a variable that falls outside of the range of the data.

Goal 3: It is not possible to separately identify G and π. To see that note that because it is πn that enters G, $G_n = \pi G_{(\pi n)}$; knowledge of G_n does not allow one to separately identify $G_{(\pi n)}$ and π. We thus cannot perform the child care subsidy policy experiment.

Parametric, Nonstructural

In this approach, one chooses a functional form for G. For example, one might choose a cumulative standard normal function in which the variables in Ω_{it}^{-} enter as a single index.

Goal 1: As in the NP-NS approach, because of the partial observability of wage offers, testing the model's prediction still requires an exclusion restriction, that is, a variable in z_{it} that is not in κ_{it}.

Goal 2: It is possible, given an estimate of G, to determine the effect on participation of a change in any of the variables not only within, but also outside, the range of the data.

Goal 3: As in the NP-NS approach, it is not possible to separately identify π from variation in n_{it} because πn_{it} enters G.

Nonparametric, Structural

In this approach, one would attempt to separately identify $U(\cdot)$, $w(\cdot)$, F from (8) without imposing auxiliary assumption about those functions. This is clearly infeasible when wages are only observed for those who work.[13]

Parametric, Structural

Although given our taxonomy, there are many possible variations on which functions to impose parametric assumptions, it is too far removed from the aims of this chapter to work through those possibilities.[14] We consider only the case in which all of the structural elements are parametric. Specifically, the structural elements are specified as follows:

$$U_{it} = c_{it} + \alpha_{it}(1 - d_{it}) \quad \text{with } \alpha_{it} = \kappa_{it}\beta_\kappa + \beta_n n_{it} + \epsilon_{it}, \tag{9}$$

$$c_{it} = y_{it} + w_{it}d_{it} - \pi n_{it}d_{it}, \tag{10}$$

$$w_{it} = z_{it}\gamma + \eta_{it}, \tag{11}$$

$$f(\epsilon_{it}, \eta_{it}) \sim N(0, \Lambda), \tag{12}$$

where $\Lambda = \begin{pmatrix} \sigma_\epsilon^2 & \cdot \\ \sigma_{\epsilon\eta} & \sigma_\eta^2 \end{pmatrix}$.[15] This specification of the model leads to a latent variable function, the difference in utilities, $U_{it}^1 - U_{it}^0$, given by

$$\begin{aligned} v_{it}^*(z_{it}, n_{it}, \kappa_{it}, \eta_{it}, \epsilon_{it}) &= z_{it}\gamma - (\pi + \beta_n)n_{it} - \kappa_{it}\beta_\kappa + \eta_{it} - \epsilon_{it} \\ &= \xi_{it}^*(\Omega_{it}^-) + \xi_{it}, \end{aligned} \tag{13}$$

where $\xi_{it} = \eta_{it} - \epsilon_{it}$, $\xi_{it}^*(\Omega_{it}^-) = z_{it}\gamma - (\pi + \beta_n)n_{it} - \kappa_{it}\beta_\kappa$ and Ω_{it}^- now consists of z_{it}, n_{it} and κ_{it}.[16]

[13] Results from Matzkin (1993) apply to the case where all wage offers are observed (regardless of participation). In that case, aside from normalizations, $w(\cdot)$, $U(\cdot)$ and the joint distribution, F are nonparametrically identified.

[14] Pagan and Ullah (1999), Chapter 7, provides a good introduction to semi-parametric estimation of discrete choice models.

[15] The unconventional assumption of normality for the wage distribution (allowing, as it does, for negative wage offers) is adopted in order to obtain a decision rule that is linear and additive in unobservables. We present a more general formulation in later sections.

[16] As we show below, the additive error (ξ_{it}) is convenient in calculating choice probabilities and is maintained for illustrative purposes. However, as we also show below, the additive structure is fragile. It is lost, for example, if the wage function takes a semi-log form or if the utility function is nonlinear in consumption. Note that the linearity and separability of consumption in the utility function implies that husband's income does not enter v_{it}^* and, thus, does not affect the participation decision.

The likelihood function, incorporating the wage information for those women who work, is

$$\begin{aligned} L(\theta; \kappa_{it_i}, z_{it_i}, n_{it_i}) &= \prod_{i=1}^{I} \Pr(d_{it_i} = 1, w_{it_i} | \Omega_{it}^{-})^{d_{it_i}} \Pr(d_{it_i} = 0 | \Omega_{it_i}^{-})^{1-d_{it_i}} \\ &= \prod_{i=1}^{I} \Pr(\xi_{it_i} \geq -\xi_{it_i}^{*}(\Omega_{it_i}^{-}), \eta_{it_i} = w_{it_i} - z_{it_i}\gamma)^{d_{it_i}} \\ &\quad \times \Pr(\xi_{it} < -\xi_{it}^{*}(\Omega_{it}^{-}))^{1-d_{it_i}}. \end{aligned} \tag{14}$$

The parameters to be estimated include β_κ, β_n, γ, π, σ_ϵ^2, σ_η^2, and $\sigma_{\epsilon\eta}$.[17] First, it is not possible to separately identify the child care cost, π, from the effect of children on the utility of not working, β_n; only $\pi + \beta_n$ is potentially identified. Joint normality is sufficient to identify the wage parameters, γ and σ_η^2, as well as $(\sigma_\eta^2 - \sigma_{\epsilon\eta})/\sigma_\xi$ (Heckman, 1979). The data on work choices identify γ/σ_ξ and β/σ_ξ. To identify σ_ξ, note that there are three possible types of variables that appear in the likelihood function, variables that appear only in z, that is, only in the wage function, variables that appear only in κ, that is, only in the value of leisure function, and variables that appear in both κ and z. Having identified the parameters of the wage function (the γ's), the identification of σ_ξ (and thus also $\sigma_{\epsilon\eta}$) requires the existence of at least one variable of the first type, that is, a variable that appears only in the wage equation.[18]

Goal 1: As in the NS approaches, there must be an exclusion restriction, in particular, a variable in z_{it} that is not in κ_{it}.

Goal 2: It is possible to determine the effect on participation of a change in any of the variables within and outside of the range of the data.

Goal 3: As noted, it is possible to identify $\pi + \beta_n$. Suppose then that a policy maker is considering implementing a child care subsidy program, where none had previously existed, in which the couple is provided a subsidy of τ dollars if the wife works when there is a young child in the household. The policy maker would want to know the impact of the program on the labor supply of women and the program's budgetary implications. With such a program, the couple's budget constraint under the child care subsidy program is

$$c_{it} = w_{it} d_{it} + y_{it} - (\pi - \tau) d_{it} n_{it}, \tag{15}$$

where $(\pi - \tau)$ is the net (of subsidy) cost of child care. With the subsidy, the probability

[17] We call $\Pr(d_{it} = 1, w_{it} | \Omega_{it}^{-})$ a probability, but it is actually a mixed probability for d and a density for w. Note that the Jacobian of the transformation from the wage density to the wage error density is one.

[18] Given the assumptions of the model, full independence of the joint error distribution with respect to observables is not necessary. See French and Taber (2011) for an extended discussion of identification of selection models.

that the woman works is

$$\Pr(d_{it} = 1|\Omega_{it}^{-}, \tau) = \Phi\left(\frac{z_{it}\gamma - \kappa_{it}\beta_{\kappa} - (\beta_n + \pi - \tau)n_{it}}{\sigma_{\xi}}\right), \tag{16}$$

where Φ is the standard normal cumulative. Given identification of $\beta_n + \pi$ from maximizing the likelihood (14), to predict the effect of the policy on participation, that is, the difference in the participation probability when τ is positive and when τ is zero, it is necessary, as seen in (16), to have identified σ_{ξ}. Government outlays on the program would be equal to the subsidy amount times the number of women with young children who work under the subsidy.

It is important to note that the policy effect is estimated without direct policy variation, i.e., we did not need to observe households in both states of the world, with and without the subsidy program. What was critical for identification was (exogenous) variation in the wage (independent of preferences). Wage variation is important in estimating the policy effect because, in the model, the child care cost is a tax on working that is isomorphic to a tax on the wage. Wage variation, independent of preferences, provides *policy-relevant* variation.

To summarize, testing the prediction that participation rises with the wage offer requires an exclusion restriction regardless of the approach. This requirement arises because of the non-observability of wage offers for those that choose not to work.[19] With regard to the second goal, the parametric approach allows extrapolation outside of the sample range of the variables whereas nonparametric approaches do not. Finally, subject to identification, the P-S approach enables the researcher to perform counterfactual exercises, subsidizing the price of child care in the example, even in the absence of variation in the child care price.[20]

3.1.2. Dynamic model

In the previously specified static model, there was no connection between the current participation decision and future utility. One way, among many, to introduce dynamic considerations is through human capital accumulation on the job. In particular, suppose that the woman's wage increases with actual work experience, h, as skills are acquired

[19] If all wage offers were observed, it would be possible to achieve all three goals without imposing parametric assumptions or structure. With respect to the policy counterfactual (goal 3), because of the subsidy acts like a wage tax, the effect of the subsidy can be calculated by comparing participation rates of women with a given wage to women with a wage augmented by πn_{it} (see Ichimura and Taber (2002) and Todd and Wolpin (2010)).

[20] Another reason for adopting the P-S estimation approach is that separating out preferences from opportunities (wage offers) helps to understand important social and economic phenomena, for example, in assessing how much of the difference in labor market outcomes of black and white women is due to differences in preferences and how much to differences in wage opportunities. Such an assessment could be useful in the design of public policies aimed at ameliorating those differences.

through learning by doing. To capture that, rewrite (11) as

$$w_{it} = z_{it}\gamma_1 + \gamma_2 h_{it} + \eta_{it}, \tag{17}$$

where $h_{it} = \sum_{\tau=1}^{\tau=t-1} d_{i\tau}$ is work experience at the start of period t. Given this specification, working in any period increases all future wage offers. Work experience, h_{it}, evolves according to

$$h_{it} = h_{i,t-1} + d_{i,t-1} \tag{18}$$

where $h_{i1} = 0$.[21] Thus, at any period t, the woman may have accumulated up to $t-1$ periods of work experience. We will be more specific about the evolution of the other state space elements when we work through the solution method below. For now, we assume only that their evolution is non-stochastic.

Normally distributed additive shocks

As in the static model, and again for presentation purposes, we assume that the preference shock (ϵ_{it}) and the wife's wage shock (η_{it}) are distributed joint normal. In addition, we assume that they are mutually serially independent and independent of observables, that is, $(f(\epsilon_{it}, \eta_{it}|\epsilon_{it-1}, \eta_{it-1,\ldots,}\epsilon_{i1}, \eta_{i1}) = f(\epsilon_{it}, \eta_{it}))$.

Assume, in this dynamic context, that the couple maximizes the expected present discounted value of remaining lifetime utility at each period starting from an initial period, $t = 1$, and ending at period T, the assumed terminal decision period.[22,23] Letting $V_t(\Omega_{it})$ be the maximum expected present discounted value of remaining lifetime utility at $t = 1, \ldots, T$ given the state space and discount factor δ,

$$V_t(\Omega_{it}) = \max_{d_{it}} E\left\{\sum_{\tau=t}^{\tau=T} \delta^{\tau-t}[U_{i\tau}^1 d_{i\tau} + U_{i\tau}^0(1-d_{i\tau})]|\Omega_{i\tau}\right\}. \tag{19}$$

[21] The assumption that the woman's initial work experience at the time marriage is zero, which is undoubtedly in many cases untrue, is made for ease of exposition. We discuss in a later section the complications introduced by accounting for the fact that work experience is accumulated prior to marriage and varies across women.

[22] The finite horizon assumption is immaterial for the points we wish to make. If the current period utility is bounded at all $t = 1, \ldots, \infty$ and the discount factor is less than one, then the solution to the infinite horizon problem can be approximated arbitrarily closely by the solution to a long but finite horizon problem. The essential difference between a finite and infinite horizon model in terms of the predictions about behavior is that in the finite horizon case there are implications for age patterns in behavior.

[23] The terminal period of the model would be at the termination of the marriage or the retirement of the wife. Accounting for divorce, even taking it to be exogenous, would unduly complicate the model. For illustrative purposes, then, we assume that the wife retires at $T+1$. The value function at $T+1$ is normalized to zero, although a more complete formulation would make the retirement decision of both spouses a choice and would, at the least, specify the determination of post-retirement income through the social security system.

The state space at t consists of the same elements as in the static model augmented to include the amount of accumulated work experience, h_{it}.

The value function ($V_t(\Omega_{it})$) can be written as the maximum over the two alternative-specific value functions, $V_t^k(\Omega_{it})$, $k \in \{0, 1\}$

$$V_t(\Omega_{it}) = \max(V_t^0(\Omega_{it}), V_t^1(\Omega_{it})), \tag{20}$$

each of which obeys the Bellman equation

$$\begin{aligned} V_t^k(\Omega_{it}) &= U_{it}^k(\Omega_{it}) + \delta E[V_{t+1}(\Omega_{i,t+1})|\Omega_{it}, d_{it} = k] \quad \text{for } t < T, \\ &= U_{iT}^k(\Omega_{iT}) \quad \text{for } t = T. \end{aligned} \tag{21}$$

The expectation in (21) is taken over the distribution of the random components of the state space at $t + 1$, $\epsilon_{i,t+1}$ and $\eta_{i,t+1}$, conditional on the state space elements at t.

The latent variable in the dynamic case is the difference in alternative-specific value functions, $V_t^1(\Omega_{it}) - V_t^0(\Omega_{it})$, namely[24]

$$\begin{aligned} v_t^*(\Omega_{it}) &= z_{it}\gamma_1 + \gamma_2 h_{it} - \pi n_{it} - \kappa_{it}\beta_\kappa - \epsilon_{it} + \eta_{it} \\ &\quad + \delta \left\{ [E[V_{t+1}(\Omega_{i,t+1})|\Omega_{it}, d_{it} = 1] \right. \\ &\quad \left. - [E[V_{t+1}(\Omega_{i,t+1})|\Omega_{it}, d_{it} = 0] \right\} \qquad (22) \\ &= \xi_{it}^*(\Omega_{it}^-) + \xi_{it}.^{25} \qquad (23) \end{aligned}$$

Comparing the latent variable functions in the dynamic (22) and static (13) cases, the only difference is the appearance in the dynamic model of the difference in the future component of the expected value functions under the two alternatives. This observation was a key insight in the development of estimation approaches for DCDP models.

To calculate these alternative-specific value functions, note first that $\Omega_{i,t+1}^-$, the observable part of the state space at $t + 1$, is fully determined by Ω_{it}^- and the choice at t, d_{it}. Thus, one needs to be able to calculate $E[V_{t+1}(\Omega_{i,t+1})|\Omega_{it}, d_{it})$ at all values of $\Omega_{i,t+1}^-$ that may be reached from the state space elements at t and a choice at t. A full solution of the dynamic programming problem consists, then, of finding $EV_\tau(\Omega_{i\tau}) = E\max[(V_\tau^0(\Omega_{i\tau}), V_\tau^1(\Omega_{i\tau}))]$ for all values of $\Omega_{i\tau}^-$ at all $\tau = 2, \ldots, T$. We denote this function by $E\max(\Omega_{it}^-)$ or $E\max_t$ for short.

[24] Given the lack of separate identification of π and β_n, we set $\beta_n = 0$ to reduce notation.

[25] Note that if preference or wage shocks were serially correlated, the observable and unobservable state variables would not generally be additively separable as in the second equality. The additive separability arises because, with serial independence, Ω_{it}^-, which does not include ϵ_{it} or η_{it} can replace Ω_{it} in the future component of the value functions,.that is $E[V_{t+1}(\Omega_{i,t+1})|\Omega_{it}, d_{it}] = E[V_{t+1}(\Omega_{i,t+1})|\Omega_{it}^-, d_{it}]$. We discuss the case of serially correlated errors below.

In the finite horizon model we are considering, the solution method is by backwards recursion. However, there are a number of additional details about the model that must first be addressed. Specifically, it is necessary to assume something about how the exogenous observable state variables evolve, that is, z_{it}, n_{it},κ_{it}.[26] For ease of presentation, to avoid having to specify the transition processes of the exogenous state variables, we assume that $z_{it} = z_i$ and $\kappa_{it} = \kappa_i$.

The number of young children, however, is obviously not constant over the life cycle. But, after the woman reaches the end of her fecund period, the evolution of n_{it} is non-stochastic.[27] To continue the example, we restrict attention to the woman's post-fecund period. Thus, during that period n_{it} is perfectly foreseen, although the future path of n_{it} at any t depends on the exact ages of the young children in the household at t.[28] Thus, the ages of existing young children at t are elements of the state space at t, Ω_{it}^-.

As seen in (21), to calculate the alternative-specific value functions at period $T-1$ for each element of $\Omega_{i,T-1}^-$, we need to calculate what we have referred to above as $E\max_T$. Using the fact that, under normality, $E(\epsilon_{iT}|\xi_{iT} < -\xi_{iT}^*(\Omega_{iT}^-)) = -\frac{\sigma_{\epsilon\xi}}{\sigma_\xi}\frac{\phi(-\xi_{iT}^*(\Omega_{iT}^-))}{\Phi(-\xi_{iT}^*(\Omega_{iT}^-))}$ and $E(\eta_{iT}|\xi_{iT} \geq -\xi_{iT}^*(\Omega_{iT}^-)) = \frac{\sigma_{\eta\xi}}{\sigma_\xi}\frac{\phi(-\xi_{iT}^*(\Omega_{iT}^-))}{1-\Phi(-\xi_{iT}^*(\Omega_{iT}^-))}$, we get

$$E\max_T = y_{iT} + (\kappa_i\beta_\kappa)\Phi(-\xi_{iT}^*(\Omega_{iT}^-)) + (z_i\gamma_1 + \gamma_2 h_{iT} - \pi n_{iT})(1 - \Phi(-\xi_{iT}^*(\Omega_{iT}^-))) + \sigma_\xi\phi(-\xi_{iT}^*(\Omega_{iT}^-)).^{29} \quad (24)$$

Note that evaluating this expression requires an integration (the normal cdf) which has no closed form; it thus must be computed numerically. The right hand side of (24) is a function of y_{iT}, z_i, κ_i, n_{iT} and h_{iT}.[30] Given a set of model parameters, the $E\max_T$ function takes on a scalar value for each element of its arguments. Noting that $h_{iT} = h_{i,T-1} + d_{i,T-1}$, and being explicit about the elements of $E\max_T$, the alternative-specific value functions at $T-1$ are (dropping the i subscript for convenience):

[26] Because of the linearity and additive separability of consumption in utility, husband's income does not affect the participation decision. We therefore do not need to specify what is known about future husband's income (see below). Again, this assumption is made so that the solution method can be illustrated most effectively.

[27] Later, we introduce stochastic fertility, allowing for the decision model to begin at the time of marriage, when we consider an extension of the model to a multinomial choice setting.

[28] Suppose we define a young child as a child under the age of six (that is, not of school age). Consider a couple who at the start of the woman's infecund period has a 3 year old child and thus for whom $n_{it} = 1$. Then, for that couple, $n_{it+1} = 1, n_{it+2} = 1$ and $n_{it+2} = \cdots = n_{iT} = 0$.

[29] This expression uses the fact that for any two random variables u and v,

$$E\max(u, v) = E(u|u > v)\Pr(u > v) + E(v|v > u)\Pr(v > u).$$

[30] Although n_{iT} would surely be zero at some point, we carry it along to emphasize its perfect foresight property.

$$V^0_{T-1}(\Omega_{T-1}) = y_{T-1} + \kappa\beta_\kappa + \epsilon_{T-1} + \delta E \max(y_T, z, \kappa, n_T, h_{T-1}), \tag{25}$$

$$V^1_{T-1}(\Omega_{T-1}) = y_{T-1} + z\gamma_1 + \gamma_2 h_{T-1} - \pi n_{T-1} + \eta_{T-1} + \delta E \max(y_T, z, \kappa, n_T, h_{T-1} + 1). \tag{26}$$

Thus,

$$\begin{aligned} v^*_{T-1}(\Omega_{i,T-1}) &= z\gamma_1 + \gamma_2 h_{T-1} - \pi n_{T-1} - \kappa\beta_\kappa - \epsilon_{T-1} + \eta_{T-1} \\ &\quad + \delta \left\{ E \max(y_{T,} z, \kappa, n_T, h_{T-1} + 1) \right. \\ &\quad \left. - E \max(y_{T,} z, \kappa, n_T, h_{T-1}) \right\} \qquad (27) \\ &= \xi^*_{T-1}(\Omega^-_{T-1}) + \xi_{T-1}. \qquad (28) \end{aligned}$$

As before, because y_T enters both $E \max(y_T, z, \kappa, n_T, h_{T-1} + 1)$ and $E \max(y_T, z, \kappa, n_T, h_{T-1})$ additively, it drops out of $\xi^*_{T-1}(\Omega^-_{T-1})$ and thus out of v^*_{T-1}.[31]

To calculate the $T-2$ alternative-specific value functions, we will need to calculate $E \max_{T-1}$. Following the development for period $T-1$,

$$\begin{aligned} E \max_{T-1} &= y_{T-1} + (\kappa\beta_\kappa + \delta E \max(y_{T-1}, z, \kappa, n_T, h_{T-1}))\Phi(-\xi^*_{T-1}(\Omega^-_{T-1})) \\ &\quad + (z\gamma_1 + \gamma_2 h_{T-1} - \pi n_{T-1} \\ &\quad + \delta E \max(y_{T-1}, z, \kappa, n_T, h_{T-1} + 1))(1 - \Phi(-\xi^*_{T-1}(\Omega^-_{T-1}))) \\ &\quad + \sigma_\xi \phi(-\xi^*_{T-1}(\Omega^-_{T-1})). \end{aligned} \tag{29}$$

The right hand side of (29) is a function of $y_{T-1}, z, \kappa, n_{T-1}, n_T$ and h_{T-1}. As with $E \max_T$, given a set of model parameters, the $E \max_{T-1}$ function takes on a scalar value for each element of its arguments. Noting that $h_{T-1} = h_{T-2} + d_{T-2}$, the alternative-specific value functions at $T-2$ and the latent variable function are given by

$$V^0_{T-2}(\Omega_{T-2}) = y_{T-2} + \kappa\beta_\kappa + \epsilon_{T-2} + \delta E \max(y_{T-1}, z, \kappa, n_{T-1}, n_T, h_{T-2}), \tag{30}$$

$$V^1_{T-2}(\Omega_{T-2}) = y_{T-2} + z\gamma_1 + \gamma_2 h_{T-2} - \pi n_{T-2} + \eta_{T-2} + \delta E \max(y_{T-1}, z, \kappa, n_{T-1}, n_T, h_{T-2} + 1), \tag{31}$$

$$\begin{aligned} v^*_{T-2}(\Omega_{T-2}) &= z\gamma_1 + \gamma_2 h_{.T-2} - \pi n_{T-2} - \kappa\beta_\kappa - \epsilon_{T-2} + \eta_{T-2} \\ &\quad + \delta \{E \max(y_{T-1}, z, \kappa, n_{T-1}, n_T, h_{T-2} + 1) \\ &\quad - E \max(y_{T-1}, z, \kappa, n_{T-1}, n_T, h_{T-2})\} \qquad (32) \\ &= \xi^*_{T-2}(\Omega^-_{T-2}) + \xi_{T-2}. \qquad (33) \end{aligned}$$

As at T, y_{T-1} drops out of $\xi^*_{T-2}(\Omega^-_{T-2})$ and thus v^*_{T-2}.

[31] In solving for the latent variable functions, we could thus set $y_t = 0$ (or any other arbitrary value) for all t.

We can continue to solve backwards in this fashion. The full solution of the dynamic programming problem is the set of $E\max_t$ functions for all t from $t = 1, \ldots, T$. These $E\max_t$ functions provide all of the information necessary to calculate the cut-off values, the $\xi_t^*(\Omega_t^-)$'s that are the inputs into the likelihood function.

Estimation of the dynamic model requires that the researcher have data on work experience, h_{it}. More generally, assume that the researcher has longitudinal data for I married couples and denote by t_{1i} and t_{Li} the first and last periods of data observed for married couple i. Note that t_{1i}need not be the first period of marriage (although it may be, subject to the marriage occurring after the woman's fecund period) and t_{Li} need not be the last (although it may be). Denoting θ as the vector of model parameters, the likelihood function is given by

$$L(\theta; \text{data}) = \prod_{i=1}^{i=I} \prod_{\tau=t_{1i}}^{\tau=t_{Li}} \Pr(d_{i\tau} = 1, w_{i\tau}|\Omega_{i\tau}^-)^{d_{i\tau}} \Pr(d_{i\tau} = 0|\Omega_{i\tau}^-)^{1-d_{i\tau}}, \tag{34}$$

where $\Pr(d_{i\tau} = 1, w_{i\tau}|\Omega_{i\tau}^-) = \Pr(\xi_{i\tau} \geq -\xi_{i\tau}^*(\Omega_{i\tau}^-), \eta_{i\tau} = w_{i\tau} - z_{i\tau}\gamma_1 - \gamma_2 h_{i\tau})$ and $\Pr(d_{i\tau} = 0|\Omega_{i\tau}^-) = 1 - \Pr(\xi_{i\tau} \geq -\xi_{i\tau}^*(\Omega_{i\tau}^-))$.[32]

Given joint normality of ϵ and η, the likelihood function is analytic, namely

$$L(\theta; \text{data}) = \prod_{i=1}^{i=I} \prod_{\tau=t_{1i}}^{\tau=t_{Li}} \left\{ \left[1 - \Phi\left(\frac{-\xi_\tau^*(\Omega_{i\tau}^-) - \rho\frac{\sigma_\xi}{\sigma_\eta}\eta_{i\tau}}{\sigma_\xi(1-\rho^2)^{\frac{1}{2}}} \right) \right] \frac{1}{\sigma_\eta}\phi\left(\frac{\eta_{i\tau}}{\sigma_\eta}\right) \right\}^{d_{i\tau}} \times \left\{ \Phi\left(\frac{-\xi_\tau^*(\Omega_{i\tau}^-)}{\sigma_\xi} \right) \right\}^{1-d_{i\tau}}, \tag{35}$$

where $\eta_{i\tau} = w_{i\tau} - z_{i\tau}\gamma_1 - \gamma_2 h_{i\tau}$ and where ρ is the correlation coefficient between ξ and η.[33] Estimation proceeds by iterating between the solution of the dynamic programming problem and the likelihood function for alternative sets of parameters. Maximum likelihood estimates are consistent, asymptotically normal and efficient.

Given the solution of the dynamic programming problem for the cut-off values, the $\xi_{it}^*(\Omega_{it}^-)$'s, the estimation of the dynamic model is in principle no different than the estimation of the static model. However, the dynamic problem introduces an additional parameter, the discount factor, δ, and additional assumptions about how households

[32] If the structure does not yield an additive (composite) error, the latent variable function becomes $v_t^*(\Omega_{it}^-, \eta_{it}, \epsilon_{it})$. Calculating the joint regions of η_{it}, ϵ_{it} that determine the probabilities that enter the likelihood function and that are used to calculate the $E\max(\Omega_{it}^-)$ function must, in that case, be done numerically. We address this more general case below.

[33] As in the static case, the Jacobian of the transformation from the density of the wage offer to the density of η is one.

forecast future unobservables.[34] The practical difference in terms of implementation is the computational effort of having to solve the dynamic programming problem in each iteration on the model parameters in maximizing the likelihood function.

Identification of the model parameters requires the same exclusion restriction as in the static case, that is, the appearance of at least one variable in the wage equation that does not affect the value of leisure. Work experience, h_{it}, would serve that role if it does not also enter into the value of leisure (κ). A heuristic argument for the identification of the discount factor can be made by noting that the difference in the future component of the expected value functions under the two alternatives in (22) is in general a nonlinear function of the state variables and depends on the same set of parameters as in the static case. Rewriting (22) as

$$v_t^*(\Omega_{it}) = z_i\gamma_1 + \gamma_2 h_{it} - \pi n_{it} - \kappa_i\beta_\kappa + \delta W_{t+1}(\Omega_{it}^-) - \epsilon_{it} + \eta_{it}, \tag{36}$$

where $W(\cdot)$ is the difference in the future component of the expected value functions, the nonlinearities in W_{t+1} that arise from the distributional and functional form assumptions may be sufficient to identify the discount factor.[35]

As in the static model, identification of the model parameters implies that all three research goals previously laid out can be met. In particular, predictions of the theory are testable, the effects on participation of changes in observables that vary in the sample are estimable and a quantitative assessment of the counterfactual child care subsidy is feasible. The effect of such a subsidy will differ from that in a static model as any effect of the subsidy on the current participation decision will be transmitted to future participation decisions through the change in work experience and thus future wages. If a surprise (permanent) subsidy were introduced at some time t, the effect of the subsidy on participation at t would require that the couple's dynamic programming problem be resolved with the subsidy from t to T and the solution compared to that without the subsidy. A pre-announced subsidy to take effect at t would require that the solution be obtained back to the period of the announcement because, given the dynamics, such a program would have effects on participation starting from the date of the announcement.[36]

Independent additive type-1 extreme value errors

When shocks are additive and come from independent type-1 extreme value distributions, as first noted by Rust (1987), the solution to the dynamic programming

[34] In the current example, couples are assumed to know the full structure of the model and to use it in forming their forecasts of future wage offers and their future preferences.

[35] It is possible that in some models additional parameters might enter W_{t+1}, say through the transition functions of state variables (see below for an example). While the same heuristic argument would apply, its validity would be less apparent.

[36] More generally, if agents have beliefs about future policies (or policy changes), such beliefs should be incorporated into the solution and estimation of the decision model.

problem and the choice probability both have closed forms, that is, they do not require a numerical integration as in the additive normal error case. The cdf of an extreme value random variable u is $\exp(-e^{-\frac{u}{\rho}})$ with mean equal to $\rho\gamma$, where γ is Euler's constant, and variance $\frac{\pi^2\rho^2}{6}$.

Under the extreme value assumption, it can be shown that for period $t = T$ (dropping the i subscript for convenience),

$$\Pr(d_T = 1|\Omega_T^-) = \exp\left(\frac{z\gamma_1 + \gamma_2 h_T - \pi n_T - \kappa\beta_\kappa}{\rho}\right) \times \left(1 + \exp\left(\frac{z\gamma_1 + \gamma_2 h_T - \pi n_T - \kappa\beta_\kappa}{\rho}\right)\right)^{-1} \tag{37}$$

$$E\max_T = \rho\left\{\gamma + \log\left[\exp\left(\frac{y_T + z\gamma_1 + \gamma_2 h_T - \pi n_T}{\rho}\right) + \exp\left(\frac{y_T + \kappa\beta_\kappa}{\rho}\right)\right]\right\} = \rho\left\{\gamma + \frac{y_T + z\gamma_1 + \gamma_2 h_T - \pi n_T}{\rho} - \log(\Pr(d_T = 1|\Omega_T^-))\right\} \tag{38}$$

and for $t < T$,

$$\Pr(d_t = 1|\Omega_t^-) = \frac{\exp(\frac{z\gamma_1+\gamma_2 h_t-\pi n_t-\kappa\beta_\kappa+\delta\{E\max_{t+1}(y_{t+1},z,\kappa,\widetilde{n}_{t+1},h_t+1)-E\max_{t+1}(y_{t+1},z,\kappa,\widetilde{n}_{t+1},h_t)\}}{\rho})}{1 + \exp(\frac{z\gamma_1+\gamma_2 h_t-\pi n_t-\kappa\beta_\kappa+\delta\{E\max_{t+1}(y_{t+1},z,\kappa,\widetilde{n}_{t+1},h_t+1)-E\max_{t+1}(y_{t+1},z,\kappa,\widetilde{n}_{t+1},h_t)\}}{\rho})} \tag{39}$$

$$E\max_t = \rho\left\{\gamma + \log\left[\exp\left(\frac{V_t^1(\Omega_t^-)}{\rho}\right) + \exp\left(\frac{V_t^0(\Omega_t^-)}{\rho}\right)\right]\right\} = \rho\left\{\gamma + \frac{y_t + z\gamma_1 + \gamma_2 h_t - \pi n_t + \delta E\max(y_{t+1}, z, \kappa, \widetilde{n}_{t+1}, h_t + 1)}{\rho} - \log(\Pr(d_t = 1|\Omega_{it}^-))\right\} \tag{40}$$

where $\widetilde{n}_{t+1}$ denotes the vector of $n_{t+1}, \ldots, n_T$ values. The solution, as in the case of normal errors, consists of calculating the $E\max_t$ functions by backwards recursion. As seen, unlike the case of normal errors, the $E\max_t$ functions and the choice probabilities have closed form solutions; their calculation does not require a numerical integration.

The extreme value assumption is, however, somewhat problematic in the labor force participation model as structured. For there to be a closed form solution to the DCDP problem, the scale parameter (ρ), and thus the error variance, must be the same for both the preference shock and the wage shock, a rather strong restriction that is unlikely to hold. The root of the problem is that the participation decision rule depends on the wage

shock. Suppose, however, that the participation model was modified so that the decision rule no longer included a wage shock. Such a modification could be accomplished in either of two ways, either by assuming that the wife's wage offer is not observed at the time that the participation decision is made or that the wage is deterministic (but varies over time and across women due to measurement error). In the former case, the wage shock is integrated out in calculating the expected utility of working. while in the latter there is no wage shock entering the decision problem. Then, by adding an independent type-1 extreme value error to the utility when the wife works, the participation decision rule will depend on the difference in two extreme value taste errors, which leads to the closed form expressions given above.

In either case, there is no longer a selection issue with respect to observed wages. Because the observed wage shock is independent of the participation decision, the wage parameters can be estimated by adding the wage density to the likelihood function for participation and any distributional assumption, such as log normality, can be assumed. In addition, as in the case of normal errors, identification of the wage parameters, along with the exclusion restriction already discussed, implies identification of the rest of the model parameters (including the scale parameter). Thus, the three research goals are achievable. Whether the model assumptions necessary to take advantage of the computational gains from adopting the extreme value distribution are warranted raises the issue how models should be judged and which model is "best," a subject we take up later in the chapter.

Unobserved state variables

We have already encountered unobserved state variables in the labor force participation model, namely the stochastic elements $(\epsilon_{it}, \eta_{it})$ in Ω_t that affect current choices. However, there may be unobserved state variables that have persistent effects through other mechanisms. Such a situation arises, for example, when the distribution of $(\epsilon_{it}, \eta_{it})$ is not independent of past shocks, that is, when $f(\epsilon_{it}, \eta_{it}|\epsilon_{it-1}, \eta_{it-1,...,}\epsilon_{i1}, \eta_{i1}) \neq f(\epsilon_{it}, \eta_{it})$.

A specific example, commonly adopted in the literature, is when shocks have a permanent-transitory structure. For reasons of tractability, it is often assumed that the permanent component takes on a discrete number of values and follows a joint multinomial distribution. Specifically,

$$\epsilon_{it} = \sum_{m^h=1}^{M} \sum_{m^w=1}^{M} \lambda_{1m^h m^w} 1(\text{type}^h = m^h, \text{type}^w = m^w) + \omega_{1it}, \tag{41}$$

$$\eta_{it} = \sum_{m^w=1}^{M} \lambda_{2m} 1(\text{type}^w = m^w) + \omega_{2it} \tag{42}$$

where there are M types each of husbands (h) and wives (w), and thus M^2 couple

types and where ω_{1it} and ω_{2it} are joint normal and iid over time.[37] Each wife's type is assumed to occur with probability π_{m^w} and each husband's type with probability π_{m^h}, with $\sum_{m^j=1}^{m^j=M} \pi_{m^j} = 1$ for $j = w, h$. A couple's type is defined by their value of $\lambda_{1m^w m^h}$, where the probability of a couple being of type (m^w, m^h) is given by $\pi_{m^w m^h}$, with $\sum_{m^w=1}^{m^w=M} \sum_{m^h=1}^{m^h=M} \pi_{m^w m^h} = 1$.[38] A couple is assumed to know their own and their spouse's type, so the state space is augmented by the husband's and wife's type. Even though types are not known to the researcher, it is convenient to add them to the state variables in what we previously defined as the observable elements of the state space, Ω_{it}^-. The reason is that, unlike the iid shocks ω_{1it} and ω_{2it}, which do not enter the $E\max_t$ functions (they are integrated out), the types do enter the $E\max_t$ functions. The dynamic programming problem must be solved for each couple's type.

The likelihood function must also be modified to account for the fact that the types are unobserved. In particular, letting $L_{(m^w,m^h)}$ be the likelihood function for a type (m^w, m^h) couple, the sample likelihood is the product over individuals of the type probability weighted sum of the type-specific likelihoods, namely

$$\prod_i L^i = \sum_{m^w=1}^{M} \sum_{m^h=1}^{M} \pi_{m^w m^h} L^i_{(m^w,m^h)}. \tag{43}$$

A second example is where the joint errors follows an ARIMA process. To illustrate, suppose that the errors follow a first-order autoregressive process, namely that $\epsilon_{i,t} = \rho_e \epsilon_{i,t-1} + \omega_{1it}$ and $\eta_{it} = \rho_\eta \eta_{i,t-1} + \omega_{2it}$, where ω_{1it} and ω_{2it} are joint normal and iid over time. Consider again the alternative-specific value functions at t, explicitly accounting for the evolution of the shocks, namely

$$\begin{aligned} V_t^k(\Omega_{it}^-, \epsilon_{it}, \eta_{it}) &= U_{it}^k(\Omega_{it}) + \delta E[V_{t+1}(\Omega_{i,t+1}^-, \epsilon_{it+1}, \eta_{it+1})|\Omega_{it}^-, \epsilon_{it}, \eta_{it}, d_{it} = k] \\ &= U_{it}^k(\Omega_{it}) + \delta E[V_{t+1}(\Omega_{i,t+1}^-, \rho_e \epsilon_{it} \\ &\quad + \omega_{1it+1}, \rho_\eta \eta_{it} + \omega_{2it+1})|\Omega_{it}^-, \epsilon_{it}, \eta_{it}, d_{it} = k], \end{aligned} \tag{44}$$

where the integration is now taken over the joint distribution of ω_{1it+1} and ω_{2it+1}. To calculate the alternative-specific value function at t, it is necessary that the $E\max_{t+1}$ function include not only $\Omega_{i,t+1}^-$, as previously specified, but also the shocks at t, ϵ_{it} and η_{it}. Thus, serial correlation augments the state space that enters the $E\max_t$ functions. The added complication is that these state space elements, unlike those we have so far considered, are continuous variables, an issue we discuss later. The likelihood function is also more complicated to calculate as it requires an integration for each couple of

[37] We ignore the possibility that the husband's type also affects his earnings because, in the model as specified, his earnings has no effect on the participation decision. In a more general specification, one would probably add this source of heterogeneity.

[38] There are obviously restrictions across the husband and wife individual type probabilities and couple type probabilities.

dimension equal to the number of observation periods (and there are two additional parameters, ρ_e and ρ_η).[39]

The existence of unobserved state variables creates also a potentially difficult estimation issue with respect to the treatment of initial conditions (Heckman, 1981). Having restricted the model to the period starting at the time the wife is no longer fecund, by that time most women will have accumulated some work experience, i.e., $h_{t_{1i}}$ will not be zero and will vary in the estimation sample. Our estimation discussion implicitly assumed that the woman's "initial" work experience, that is, work experience at t_{1i}, could be treated as exogenous, that is, as uncorrelated with the stochastic elements of the future participation decisions. When there are unobserved initial state variables, permanent types or serially correlated shocks, this assumption is unlikely to hold.

Although we have not specified the labor force participation model governing decisions prior to this period, to avoid accounting for fertility decisions, it is reasonable to suppose that women who worked more while they were of childbearing ages come from a different type distribution than women who worked less, or, in the case in which there are serially correlated shocks, women with greater work experience during the childbearing period may have experienced shocks (to wages or preferences) that are not uncorrelated with those that arise after. Put differently, it would seem much more reasonable to assume that the same model governs the participation decision during pre- and post-childbearing ages than to assume that there are two different models in which decisions across those periods are stochastically independent (conditional on observables).

There are several possible solutions to the initial conditions problem. Suppose for the sake of exposition, though unrealistically, that all women begin marriage with zero work experience.[40] At the time of marriage, in the case of permanent unobserved heterogeneity, the couple is assumed to be "endowed" with a given set of preferences. A couple who intrinsically places a low value on the wife's leisure will be more likely to choose to have the wife work and thus accumulate work experience. Such women will have accumulated more work experience upon reaching the end of their childbearing years than women in marriages where the wife's value of leisure is intrinsically greater. Thus, when the end of the childbearing years are reached, there will be a correlation between the accumulated work experience of wives and the preference endowment, or type, of couples.

Suppose that participation decisions during the childbearing years were governed by the same behavioral model (modified to account for fertility) as those during the infecund observation period. In particular, suppose that given a couple's type, all shocks (the ω's in (41) and (42)) are iid. In that case, work experience can be taken as exogenous

[39] We could combine the permanent-transitory scheme with the AR(1) scheme by allowing the ω_{1it} and ω_{2it} shocks in (41) and (42) to be AR(1).

[40] Alternatively as noted, we could assume, unrealistically as well, that the experience that women have at the start of marriage is exogenous with respect to future participation decisions.

conditional on a couple's type. To condition the likelihood (43) on initial experience, we specify a type probability function conditional on work experience at the beginning of the infecund period. Specifically, we would replace $\pi_{m^w m^h}$, taken to be scalar parameters in the likelihood function (43), with the type probability function $\pi_{m^w m^h}(h_{t_{1i}})$, where, as previously defined, t_{1i} is the first (post-childbearing) period observed for couple i.[41]

The type probability function can itself be derived using Bayes' rule starting from the true initial decision period (taken to be the start of marriage in this example). Specifically, denoting the couple's endowment pair (m^w, m^h) as "type" and dropping the i subscript, because

$$L(d_{t_1-1,\ldots,}d_2, d_1|\text{type})\Pr(\text{type}) = \Pr(\text{type}|d_{t_1-1,\ldots,}d_2, d_1)\Pr(d_{t_1-1,\ldots,}d_2, d_1) \quad (45)$$

$$= \Pr(\text{type}|h_{t_1}) \times \sum_{\text{type}} L(d_{t_1-1,\ldots,}d_2, d_1|\text{type})\Pr(\text{type}), \quad (46)$$

the type probability function is

$$\pi_{m^w m^h}(h_{t_{1i}}) = \Pr(\text{type}|h_{t_1}) = \frac{L(d_{t_1-1,\ldots,}d_2, d_1|\text{type})\Pr(\text{type})}{\sum_{\text{type}} L(d_{t_1-1,\ldots,}d_2, d_1|\text{type})\Pr(\text{type})}. \quad (47)$$

Estimating the type probability function $\pi_{m^w m^h}(h_{t_{1i}})$ as a nonparametric function of $h_{t_{1i}}$ provides an "exact" solution (subject to sampling error) to the initial conditions problem, yielding type probabilities for each level of experience that would be the same as those obtained if we had solved and estimated the model back to the true initial period and explicitly used (47). Alternatively, because the type probabilities must also be conditioned on all other exogenous state variables (the z and κ variables), perhaps making nonparametric estimation infeasible, estimating a flexible functional form would provide an "approximate" solution.

If the shocks are serially correlated, work experience at the start of the infecund period is correlated with future choices not only because it affects future wages, but also because of the correlation of stochastic shocks across fecund and infecund periods. In that case, as suggested by Heckman (1981) in a nonstructural setting, we would need to have data on exogenous initial conditions at the time of the true initial period (taken here to be the start of marriage), when the labor supply decision process is assumed to begin. Given that, we can specify a density for work experience as a function of those exogenous initial conditions at the start of marriage and incorporate it in the likelihood function.[42]

[41] We would also need to include any other initial conditions that affect wage offers (z_i's) or preferences (κ_i's), for example, completed schooling.

[42] If there is both unobserved permanent heterogeneity and serial correlation, and letting Ω_0 be the exogenous initial conditions at the time of marriage, then in the likelihood function (43), $\pi_{m^w m^h}$ would be replaced with $\pi_{m^w m^h}(h_{t_1}, z, \kappa) \times \Pr(h_{t_1}|\Omega_0, z, \kappa)$. Note that Ω_0 must contain a variable other than z and κ in order to identify the effect of h_{t_1} on a couple's type.

The curse of dimensionality

As we have seen, the solution of the dynamic programming problem required that the $E\max_t$ functions be calculated for each point in the state space. If z and κ take on only a finite number of discrete values (e.g., years of schooling, number of children), as does h_t, the solution method simply involves solving for the $E\max_t$ functions at each point in the state space. However, if either z or κ contains a continuous variable (or if the shocks follow an ARIMA process, as already discussed), the dimensionality of the problem is infinite and one obviously cannot solve the dynamic programming problem at every state point. Furthermore, one could imagine making the model more complex in ways that would increase the number of state variables and hence the size of the state space, for example, by letting the vector of taste shifters κ include not just number of children but the number of children in different age ranges. In general, in a finite state space problem, the size of the state space grows exponentially with the number of state variables. This is the so-called curse of dimensionality, first associated with Bellman (1957).

Estimation requires that the dynamic programming problem be solved many times—once for each trial parameter vector that is considered in the search for the maximum of the likelihood function (and perhaps at many nearby parameter vectors, to obtain gradients used in a search algorithm). This means that an actual estimation problem will typically involve solving the DP problem thousands of times. Thus, from a practical perspective, it is necessary that one be able to obtain a solution rather quickly for estimation to be feasible. In practice, there are two main ways to do this. One is just to keep the model simple so that the state space is small. But, this precludes studying many interesting problems in which there are a large set of choices that are likely to be interrelated (for example, choices of fertility, labor supply, schooling, marriage and welfare participation).

A second approach, which a number of researchers have pursued in recent years, is to abandon "exact" solutions to DP problems in favor of approximate solutions that can be obtained with greatly reduced computational time. There are three main approximate solution methods that have been discussed in the literature:[43]

1. Discretization: This approach is applicable when the state space is large due to the presence of continuous state variables. The idea is straightforward: simply discretize the continuous variables and solve for the $E\max_t$ functions only on the grid of discretized values. To implement this method one must either (i) modify the law of motion for the state variables so they stay on the discrete grid (e.g., one might work with a discrete AR(1) process) or (ii) employ a method to interpolate between grid points. Clearly, the finer the discretization, the closer the approximation will be to the exact solution. Discretization does not formally break the curse of dimensionality because the time

[43] Note that we do not discuss methods like Hotz and Miller (1993) here. They propose a method to circumvent having to obtain a full solution of the DP problem while still obtaining parameter estimates, not a method for solving the DP problem (see below).

required to compute an approximate solution still increases exponentially as the number of state variables increases. But it can be an effective way to reduce computation time in a model with a given number of state variables.

2. Approximation and interpolation of the $E\max_t$ functions: This approach was originally proposed by Bellman et al. (1963) and extended to the type of models generally of interest to labor economists by Keane and Wolpin (1994). It is applicable when the state space is large either due the presence of continuous state variables or because there are a large number of discrete state variables (or both). In this approach the $E\max_t$ functions are evaluated at a subset of the state points and some method of interpolation is used to evaluate $E\max_t$ at other values of the state space. This approach requires that the $E\max_t$ interpolating functions be specified parametrically. For example, they might be specified as some regression function in the state space elements or as some other approximating function such as a spline. Using the estimated values of the $E\max_t$ rather than the true values is akin to having a nonlinear model with specification error. The degree of approximation error is, however, subject to control. In a Monte Carlo study, Keane and Wolpin (1994) provide evidence on the effect of this approximation error on the bias of the estimated model parameters under alternative interpolating functions and numbers of state points. Intuitively, as the subset of the state points that are chosen is enlarged and the dimension of the approximating function is increased, the approximation will converge to the true solution.[44]

As with discretization, the approximation/interpolation method does not formally break the curse of dimensionality, except in special cases. This is because the curse of dimensionality applies to polynomial approximation (see Rust (1997)). As the number of state variables grows larger, the computation time needed to attain a given accuracy in a polynomial approximation to the Emax function grows exponentially.[45] Despite this, the Keane and Wolpin (1994) approach (as well as some closely related variants) has proven to be a useful way to reduce computation time in models with large state spaces, and it has been widely applied in recent years. Rather than describe the method in detail here, we will illustrate the method later in a specific application.

3. Randomization: This approach was developed by Rust (1997). It is applicable when the state space is large due the presence of continuous state variables, but it requires that choice variables be discrete and that state variables be continuous. It also imposes important constraints on how the state variables may evolve over time. Specifically, Rust (1997) shows that solving a random Bellman equation can break the curse of dimensionality in the case of DCDP models in which the state space is continuous

[44] There is no formal proof of this proposition, though, as noted, Keane and Wolpin (1994) provide Monte Carlo evidence for a particular model that supports the intuition.

[45] Geweke and Keane (2001) give an example where the curse of dimensionality is broken. This is when the Emax can be expressed as a function of the expected value of each alternative. (That is, these expected values are a sufficient statistic for all the state variables that determine them.) The size of this set of variables remains fixed at J, where J is the number of alternatives, even as the state space grows larger.

and evolves stochastically, conditional on the alternative chosen. Note that because work experience is discrete and evolves deterministically in the labor force participation model presented above, this method does not strictly apply. But, suppose instead that we modeled work experience as a continuous random variable with density function $p(h_{t+1}|h_t, d_t) = p(h_t + jI(d_t = 1) - jI(d_t = 0)|h_t, d_t)$ where j is random variable indicating the extent to which working probabilistically augments work experience or not working depletes effective work experience (due to depreciation of skills). The random Bellman equation (ignoring z and κ), the analog of (20), is in that case given by

$$\widehat{V}_{Mt}(h_t) = \max_{d_t} \left[U_t^{d_t}(h_t) + \frac{\delta}{M} \sum_{m=1}^{M} \widehat{V}_{M,t+1}(h_{t+1,m}|h_t, d_t) p(h_{t+1,m}|h_t, d_t) \right], \quad (48)$$

where $[h_{t+1,1}, \ldots, h_{t+1,M}] = [h_1, \ldots, h_M]$ are M randomly drawn state space elements. The approximate value function $\widehat{V}_{Mt}(h_t)$ converges to $V_t(h_t)$ as $M \to \infty$ at a $\sqrt{M}$ rate. Notice that this is still true if (h_t) is a vector of state variables, regardless of the dimension of the vector. Thus, the curse of dimensionality is broken here, exactly analogously to the way that simulation breaks the curse of dimensionality in approximation of multivariate integrals (while discretization methods and quadrature do not).[46]

The above approach only delivers a solution for the value functions on the grid $[h_1, \ldots, h_M]$. But forming a likelihood will typically require calculating value functions at other points. A key point is that $\widehat{V}_{Mt}(h_t)$ is, in Rust's terminology, self-approximating. Suppose we wish to construct the alternative specific value function $\widehat{V}_{M_t}^{d_t}(h_t)$ at a point h_t that is not part of the grid $[h_1, \ldots, h_M]$. Then we simply form:

$$\widehat{V}_{M_t}^{d_t}(h_t) = U_t^{d_t}(d_t) + \delta \sum_{m=1}^{M} \widehat{V}_{M_t}(h_m) \frac{p(h_m|h_t, d_t)}{\sum_{k=1}^{M} p(h_k|h_t, d_t)}. \quad (49)$$

Notice that, because any state space element at $t + 1$ can be reached from any element at t with some probability given by $p(\cdot|h_t, d_t)$, the value function at t can be calculated from (49) at any element of the state space at t. In contrast to the methods of approximation described above, the value function does not need to be interpolated using an auxiliary interpolating function.[47] This "self-interpolating" feature of the random Bellman equation is also crucial for breaking the curse of dimensionality (which, as noted above, plagues interpolation methods).

[46] Technically this is not quite enough, as convergence must be uniform and not just pointwise.

[47] Because h_t is now an unobserved component of the state space, estimation of $p(h_{t+1|}|h_{t,}, d_t)$ must be carried out jointly. This would require a distributional assumption for p and raises issues of the separate identification of p and of the effect of h_t on wages.

Of course, the fact that the randomization method breaks the curse of dimensionality does not mean it will outperform other methods in specific problems. That the method breaks the curse of dimensionality is a statement about its behavior under the hypothetical scenario of expanding the number of state variables. For any given application with a given number of state variables, it is an empirical question whether a method based on discretization, approximation/interpolation or randomization will produce a more accurate approximation in given computation time.[48] Obviously more work is needed on comparing alternative approaches.[49]

3.2. The multinomial dynamic discrete choice problem

The structure of the labor force decision problem described above was kept simple to provide an accessible introduction to the DCDP methodology. In this section, we extend that model to allow for:

(i) additional choices;
(ii) nonadditive errors;
(iii) general functional forms and distributional assumptions.

The binary choice problem considers two mutually exclusive alternatives, the multinomial problem more than two. The treatment of static multinomial choice problems is standard. The dynamic analog to the static multinomial choice problem is conceptually no different than in the binary case. In terms of its representation, it does no injustice to simply allow the number of mutually exclusive alternatives, and thus the number of alternative-specific value functions in (21), to be greater than two. Analogously, if there are $K > 2$ mutually exclusive alternatives, there will be $K - 1$ latent variable functions (relative to one of the alternatives, arbitrarily chosen). The static multinomial choice problem raises computational issues with respect to the calculation of the likelihood function. Having to solve the dynamic multinomial choice problem, that is, for the $E\max[V_t^0(\Omega_{it}), V_t^1(\Omega_{it}), \ldots, V_t^K(\Omega_{it})]$ function that enters the multinomial version of (21) at all values of Ω_{it}^- and at all t, adds significant computational burden.

For concreteness, we consider the extension of DCDP models to the case with multiple discrete alternatives by augmenting the dynamic labor force participation model to include a fertility decision in each period so that the model can be extended to childbearing ages. In addition, to capture the intensive work margin, we allow the couple to choose among four labor force alternatives for the wife. We also drop the assumption that errors are additive and normal. In particular, in the binary model we assumed, rather unconventionally, that the wage has an additive error in levels. The usual specification (based on both human capital theory and on empirical fit) is that the *log* wage has an

[48] This is analogous to the fact that the asymptotic properties of competing estimators (under the hypothetical scenario of increasing sample size) do not reveal which will perform best given finite samples.

[49] Stinebrickner (2000) compares several approximation methods in the context of a DCDP model with serially correlated shocks.

additive error.[50] Although it is necessary to impose functional form and distributional assumptions to solve and estimate DCDP models, it is not necessary to do so to describe solution and estimation procedures. We therefore do not impose such assumptions, reflecting the fact that the researcher is essentially unconstrained in the choice of parametric and distributional assumptions (subject to identification considerations).

The following example also illustrates the interplay between model development and data. The development of a model requires that the researcher decide on the choice set, on the structural elements of the model and on the arguments of those structural elements. In an ideal world, a researcher, based on prior knowledge, would choose a model, estimate it and provide a means to validate it. However, in part because there are only a few data sets on which to do independent validations and in part because it is not possible to foresee where models will fail to fit important features of data, the process by which DCDP models are developed and empirically implemented involves a process of iterating among the activities of model specification, estimation and model validation (for example, checking model fit). Any empirical researcher will recognize this procedure regardless of whether the estimation approach is structural or nonstructural.

A researcher who wished to study the relationship between fertility and labor supply of married women would likely have in mind some notion of a model, and, in that context, begin by exploring the data. A reasonable first step would be to estimate regressions of participation and fertility as functions of "trial" state variables, interpreted as approximations to the decision rules in a DCDP model.[51] As an example, consider a sample of white married women (in their first marriage) taken from the 1979-2004 rounds of the NLSY79. Ages at marriage range from 18 to 43, with 3/4ths of these first marriages occurring before the age of 27. We adopt, as is common in labor supply models, a discrete decision period to be a year.[52] The participation measure consists of four mutually exclusive and exhaustive alternatives, working less than 500 hours during a calendar year ($d_{it}^0 = 1$), working between 500 and 1499 hours ($d_{it}^1 = 1$), working between 1500 and 2499 hours ($d_{it}^2 = 1$) and working more than 2500 hours ($d_{it}^3 = 1$).[53]

[50] The Ben-Porath (1967) model of human capital accumulation leads to a semi-log form and Heckman and Polachek (1974) show using a Box-Cox transformation that a semi-log form is not rejected by the data.

[51] Approximations to DCDP model decision rules were first discussed in Heckman (1981) and Wolpin (1984). For an empirical application in the labor economics literature, see Keane and Wolpin (2001).

[52] In theory, the period length should correspond to the frequency of decision-making, which, in principle, may differ among choice variables. Like the specification of the model structure (including assumptions about expectations formation and optimization), the discrete time framework is adopted as an approximation. A continuous time framework would be more general, but would require assumptions about the joint process generating decision times for the choice variables.

[53] To the extent that variations in hours worked within those categories represents differences in the choice of optimal hours, the discretization of hours induces measurement error. In the data, the mean and standard deviation of hours based on the categorization (where the categories are assigned 0, 1000, 2000 and 3000 hours) are almost identical to that based on actual annual hours worked. The standard deviation of hours within the categories is 145, 286, 224 and 429.

The fertility measure is the dichotomous variable indicating whether or not the woman had a birth during the calendar year. The approximate decision rule for participation is estimated by an ordered probit and the fertility decision rule by a binary probit. The variables included in these approximate decision rules, corresponding to the original taxonomy in section II, are $\widetilde{D}_{it}$ = {total hours worked up to t, hours worked in $t-1$, whether a child was born in $t-1$, number of children born between $t-2$ and $t-5$, number of children ever born, t (years of marriage up to t)} and $\widetilde{X}_{ijt}$= {age of wife, age of spouse, schooling of wife, schooling of spouse}. Consistent with any DCDP model, the same state variables enter the approximate decision rules for participation and for fertility. As seen in Table 1, the state variables appear to be related to both decision variables and in reasonable ways.[54]

Suppose the researcher is satisfied that the state variables included in the approximate decision rules should be included in the DCDP model. The researcher, however, has to make a choice as to where in the set of structural relationships the specific state variables should appear: the utility function, the market wage function, the husband's earnings function and/or the budget constraint. The researcher also must decide about whether and where to include unobserved heterogeneity and/or serially correlated errors. Some of these decisions will be governed by computational considerations. Partly because of that and partly to avoid overfitting, researchers tend to begin with parsimonious specifications in terms of the size of the state space. The "final" specification evolves through the iterative process described above.

As an example, let the married couple's per-period utility flow include consumption (c_{it}), a per-period disutility from each working alternative and a per-period utility flow from the stock of children (N_{it}). The stock of children includes a newborn, that is a child born at the beginning of period t ($n_{it}=1$). Thus,

$$U_{it} = U(c_{it}, d_{it}^1, d_{it}^2, d_{it}^3, N_{it}; \epsilon_{it}^1, \epsilon_{it}^2, \epsilon_{it}^3, \epsilon_{it}^n), \tag{50}$$

where the $\epsilon_{it}^1, \epsilon_{it}^2, \epsilon_{it}^3$, and ϵ_{it}^n are time-varying preference shocks associated with each of the four choices that are assumed to be mutually serially uncorrelated. Allowing for unobserved heterogeneity, the type specification is (following (41))

$$\epsilon_{it}^j = \sum_{m^h=1}^{M} \sum_{m^w=1}^{M} \lambda_{1m}^j 1(\text{type}^h = m^h, \text{type}^w = m^w) + \omega_{1it}^j, \quad j = 1, 2, 3, n, \tag{51}$$

where the ω^j's are mutually serially independent shocks.

[54] We have not, however, in this exploratory stage allowed for serially correlated unobservables either through permanent unobserved heterogeneity or serially correlated shocks.

Table 1 Employment and fertility of married (white) women: NLSY79

	Employment hours (ordered probit)[a]	**Fertility (probit)[b]**
Work experience (hours)	4.09 E−05	8.32E−06
	(3.22E−06)[c]	(4.33E−06)
Hours $(t-1)=1$	1.04	−0.047
	(0.042)	(0.051)
Hours $(t-1)=2$	1.90	−0.126
	(0.049)	(0.051)
Hours $(t-1)=3$	3.16	−0.222
	(0.110)	(0.089)
Age	−0.075	0.211
	(0.008)	(0.035)
Age squared	—	(−0.004)
		(0.0005)
Birth $(t-1)$	−0.497	−0.320
	(0.047)	(0.778)
Births $(t-1$ to $t-5)$	−0.349	0.448
	(0.031)	(0.054)
Total births	0.099	−0.337
	(0.028)	(0.061)
Schooling	0.077	0.004
	(0.009)	(0.011)
Age of spouse	0.007	−0.016
	(0.004)	(0.004)
Schooling of spouse	−0.036	0.021
	(0.007)	(0.010)
Marital duration	−0.025	−0.015
	(0.006)	(0.008)
Constant	—	−3.41
		(0.497)
Cut point	−0.888	—
	(0.171)	
Cut point	0.076	—
	(0.172)	
Cut point	2.48	—
	(0.175)	
Pseudo R2	.295	.094

[a] 8183 person-period observations.
[b] 8786 person-period observations.
[c] Robust standard errors in parenthesis.

The household budget constraint incorporates a cost of avoiding a birth (contraceptive costs, b_0), which, for biological reasons, will be a function of the wife's age (her age at marriage, a_0^w, plus the duration of marriage, t) and (child) age-specific monetary costs of supplying children with consumption goods (b_{1k}) and with child care if the woman works (b_{2k} per work hour). Household income is the sum of husband's earnings (y_{it}) and wife's earnings, the product of an hourly wage (w_{it}) and hours worked (1000 hours if $d_{it}^1 = 1$, 2000 hours if d_{it}^2, 3000 hours $d_{it}^3 = 1$). Specifically, the budget constraint is

$$c_{it} = y_{it} + w_{it}(1000d_{it}^1 + 2000d_{it}^2 + 3000d_{it}^3) - b_0(a_0^w + t)(1 - n_{it}) - \sum_{k=1}^{K} b_{1k} N_{kit}$$
$$- \sum_{k=1}^{K} b_{2k} N_{kit}(1000d_{it}^1 + 2000d_{it}^2 + 3000d_{it}^3) \tag{52}$$

where N_{kit} are the number of children in K different age classes, e.g., 0-1, 2-5, etc.[55] To simplify, we do not allow for uncertainty about births. A couple can choose to have a birth (with probability one) and thus not pay the contraceptive cost or choose not to have a birth (with probability one) and pay the avoidance cost.[56]

The wife's Ben Porath-Griliches wage offer function depends on her level of human capital, Ψ_{it}, which is assumed to be a function of the wife's completed schooling (S_i^w), assumed fixed after marriage, the wife's work experience, that is, the number of hours worked up to t, E_{it}, and on the number of hours worked in the previous period:

$$\log w_{it} = \sum_{j=1}^{3} \log r^j d_{it}^j + \log \Psi_{it}(S_i^w, E_{it}, d_{it-1}^1, d_{it-1}^2, d_{it-1}^3; \eta_{it}^w), \tag{53}$$

$$\eta_{it}^w = \sum_{m^w=1}^{M} \lambda_{2m^w} 1(\text{type}^w = m^w) + \omega_{2it}^w, \tag{54}$$

where the r^j are (assumed to be time-invariant) competitively determined skill rental prices that may differ by hours worked and η_{it}^w is a time varying shock to the wife's human capital following a permanent (discrete type)-transitory scheme.[57] Husband's earnings depends on his human capital according to:

55 The constant term in the contraceptive cost function, say b_{00} cannot be separately identified from b_{11}, that is, the goods cost of a newborn (a child age 0-1). Note that $N_{1it} = n_{it}$.

56 One could instead allow for a choice of whether to contracept or not with pregnancy being an uncertain outcome. We ignore this extension for ease of presentation.

57 In Ben-Porath's (1967) model of the production of human capital, an individual's wage was given by the product of a human capital per-unit rental price times the individual's human capital stock. Griliches (1977) operationalized the human capital production function as depending on arguments such as schooling, work experience and ability.

$$\log y_{it} = \log r^h + \log \Psi_{it}^h(S_i^h, a_t^h; \eta_{it}^h), \tag{55}$$

$$\eta_{it}^h = \sum_{m^h=1}^{M} \lambda_{2m^h}^h 1(\text{type}^h = m^h) + \omega_{2it}^y \tag{56}$$

where S_i^h is the husband's schooling and $a_t^h = a_0^h + t$ is his age at t (his age at marriage plus t).[58]

The time-varying state variables, the stock of children (older than one) of different ages, the total stock of children and work experience, evolve according to:

$$N_{2it} = \sum_{j=t-1}^{t-5} n_{ij}; \qquad N_{3it} = \sum_{j=t-6}^{t-17} n_{ij}; \qquad N_{it} = N_{it-1} + n_{it}, \tag{57}$$

$$E_{it} = E_{it-1} + 1000d_{it-1}^1 + 2000d_{it-1}^2 + 3000d_{it-1}^3. \tag{58}$$

The state variables in Ω_t^-, augmented to include type, consist of the stock of children (older than one) of different ages, the wife's work experience and previous period work status, the husband's and wife's age at marriage, the husband and wife's schooling levels and the couple's type. The choice set during periods when the wife is fecund, assumed to have a known terminal period (t_m), consists of the four work alternatives plus the decision of whether or not to have a child. There are thus eight mutually exclusive choices, given by $d_{it}^{hn} = \{d_{it}^{00}, d_{it}^{10}, d_{it}^{20}, d_{it}^{30}, d_{it}^{01}, d_{it}^{11}, d_{it}^{21}, d_{it}^{31} : t = 1, \ldots, t_m - 1\}$, where the first superscript refers to the work choice ($h = \{0, 1, 2, 3\}$) and the second to the fertility choice ($n = \{0, 1\}$).[59] When the wife is no longer fecund, $n_{it} = 0$ and the choice set consists only of the four mutually exclusive alternatives, $d_{it}^{hn} = \{d_{it}^{00}, d_{it}^{10}, d_{it}^{20}, d_{it}^{30} : t = t_m, \ldots, T\}$.

The objective function of the couple is, as in the binary case, to choose the mutually exclusive alternative at each t that maximizes the remaining expected discounted value of the couple's lifetime utility. Defining U_{it}^{hn} to be the contemporaneous utility flow for the work and fertility choices, the alternative-specific value functions for the multinomial choice problem are

$$\begin{aligned} V_t^{hn}(\Omega_{it}) &= U_{it}^{hn}(\Omega_{it}) + \delta E[V_{t+1}(\Omega_{i,t+1})|\Omega_{it}^-, d_{it}^{hn}] \quad \text{for } t < T, \\ &= U_{iT}^{hn}(\Omega_{iT}) \quad \text{for } t = T, \end{aligned} \tag{59}$$

where, letting $\widetilde{V_t^{hn}}$ be the vector of alternative specific value functions relevant at

[58] Husband's are assumed to work full-time, which implies that, given schooling, age and work experience are isomorphic.

[59] For convenience, h is in 1000 hour units.

period t,

$$V_t(\Omega_{it}) = \max(\widetilde{V_t^{hn}}(\Omega_{it})), \tag{60}$$

and where the expectation in (59) is taken over the joint distribution of the preference and income shocks, $f(\omega_{1t}^1, \omega_{1t}^2, \omega_{1t}^3, \omega_{1t}^n, \omega_{2t}^w, \omega_{2t}^y)$.[60] The ω's may have a general contemporaneous correlation structure, but, as noted, are mutually serially independent.

The model is solved by backwards recursion. The solution requires, as in the binary case, that the $E\max_t$ function be calculated at each state point and for all t. In the model as it is now specified, the $E\max_t$ function is a six-variate integral (over the preference shocks, the wife's wage shock and the husband's earnings shock). The state space at t consists of all feasible values of $E_{it}, d_{it-1}^1, d_{it-1}^2, d_{it-1}^3, S^w, S^h, N_{it-1}, N_{kit}$ $(k = 2, 3)$, $a_0^h, a_0^w, \text{type}^h, \text{type}^w$. Notice that all of the state variables are discrete and the dimension of the state space is therefore finite. However, the state space, though finite, is huge. The reason is that to keep track of the number of children in each of the three age groups, it is necessary to keep track of the complete sequence of births. If a woman has say 30 fecund periods, the number of possible birth sequences is $2^{30} = 1{,}073{,}700{,}000$. Even without multiplying by the dimension of the other state variables, full solution of the dynamic programming problem is infeasible, leaving aside the iterative process necessary for estimation.

It is thus necessary to use an approximation method, among those previously discussed, for solving the dynamic programming problem, that is, for solving for the $E\max_t$ functions. As an illustration, we present an interpolation method based on regression. To see how it works, consider first the calculation of the $E\max_T$ for any given state space element. At T the woman is no longer fecund, so we need to calculate

$$E\max_T = E_{T-1}\max(U_T^{00}(\widetilde{\omega}), U_T^{10}(\widetilde{\omega}), U_T^{20}(\widetilde{\omega}), U_T^{30}(\widetilde{\omega})), \tag{61}$$

where $\widetilde{\omega}$ is the six-tuple vector of shocks. Although this expression is a six-variate integration, at most four of the shocks actually affect U_T^{hn} for any given h, n choice. Given the lack of a closed form expression, $E\max_T$ must be calculated numerically. A straightforward method is Monte Carlo integration. Letting $\widetilde{\omega}_d$ be the d^{th} random draw, $d = 1, \ldots, D$, from the joint distribution, $f(\omega_1^1, \omega_1^2, \omega_1^3, \omega_1^N, \omega_2^w, \omega_2^h)$, an estimate of $E\max_T$ at say the kth value of the state space in Ω_T^-, Ω_{Tk}^-, is

$$\begin{aligned}\widehat{E\max_{Tk}} = \frac{1}{D}\sum_{d=1}^{D} \max[&U_T^{00}(\widetilde{\omega_d}; \Omega_{Tk}^-), U_T^{10}(\widetilde{\omega_d}; \Omega_{Tk}^-),\\ &U_T^{20}(\widetilde{\omega_d}; \Omega_{Tk}^-), U_T^{30}(\widetilde{\omega_d}; \Omega_{Tk}^-)].\end{aligned} \tag{62}$$

[60] U_t^{hn} is obtained after substituting for the wife's wage and the husband's earnings in the budget constraint and then substituting for consumption in the utility function.

Given the infeasibility of calculating $\widehat{E\max_T}$ at all points in the state space, suppose one randomly draws K_T state points (without replacement) and calculates the $\widehat{E\max_T}$ function for those K_T state space elements according to (62). We can treat these K_T values of $\widehat{E\max_T}$ as a vector of dependent variables in an interpolating regression

$$\widehat{E\max_{Tk}} = g_T(\Omega_{Tk}^-; \gamma_T) + \varsigma_{Tk}, \tag{63}$$

where γ_T is a time T vector of regression coefficients and $g_T(\cdot;\cdot)$ is a flexible function of state variables.[61] With this interpolating function in hand, estimates of the $E\max_T$ function can be obtained at any state point in the set Ω_T^-.

Given $\widehat{E\max_T}$, we can similarly calculate V_{T-1}^{hn} at a subset of the state points in Ω_{T-1}^-. Using the D draws from $f(\widetilde{\omega})$, the estimate of $E\max_{T-1}$ at the kth state space element is

$$\begin{aligned}\widehat{E\max_{T-1,k}} = \frac{1}{D}\sum_{d=1}^{D} \max[&V_{T-1}^{00}(\widetilde{\omega_d}; \Omega_{T-1,k}^-), V_{T-1}^{10}(\widetilde{\omega_d}; \Omega_{T-1,k}^-),\\ &V_{T-1}^{20}(\widetilde{\omega_d}; \Omega_{T-1\,,k}^-), V_{T-1}^{30}(\widetilde{\omega_d}; \Omega_{T-1,k}^-)],\end{aligned} \tag{64}$$

where V_{T-1}^{hn} is given by (59). Using the $\widehat{E\max_{T-1,k}}$ calculated for K_{T-1} randomly drawn state points from Ω_{T-1}^- as the dependent variables in the interpolating function,

$$\widehat{E\max_{T-1,k}} = g_{T-1}(\Omega_{T-1,k}^-; \gamma_{T-1}) + \varsigma_{T-1}, \tag{65}$$

provides estimated values for the $E\max_{T-1}$ function at any state point in the set Ω_{T-1}^-.[62] Continuing this procedure, we can obtain the interpolating functions for all of the $\widehat{E\max_t}$ functions for all t from t_m (the age at which the woman becomes infertile) through T, that is, $g_T, g_{T-1}, \ldots, g_{t_m}$.

At $t = t_m - 1$, the choice set now includes the birth of a child. All of the $E\max_t$ functions from $t = 1$ to $t_m - 1$ require numerical integrations over the eight mutually exclusive choices based on the joint error distribution $f(\widetilde{\omega})$. At any t within the fecund period, at the kth state point,

$$\begin{aligned}\widehat{E\max_{tk}} = \frac{1}{D}\sum_{d=1}^{D} \max[&V_t^{00}(\widetilde{\omega_d}; \Omega_{tk}^-), V_t^{10}(\widetilde{\omega_d}; \Omega_{tk}^-), V_t^{20}(\widetilde{\omega_d}; \Omega_{tk}^-), V_t^{30}(\widetilde{\omega_d}; \Omega_{tk}^-),\\ &V_t^{01}(\widetilde{\omega_d}; \Omega_{tk}^-), V_t^{11}(\widetilde{\omega_d}; \Omega_{tk}^-), U_t^{21}(\widetilde{\omega_d}; \Omega_{tk}^-), U_t^{31}(\widetilde{\omega_d}; \Omega_{tk}^-)].\end{aligned} \tag{66}$$

61 Keane and Wolpin (1994) discuss various specifications of the regression function.

62 In the labor force participation model, the total number of potential state points increases in t as feasible work experience and numbers of children increase. A researcher might, as the notation indicates, vary the number of randomly drawn state points with t.

Again taking K_t random draws from the state space at t, we can generate interpolating functions:[63]

$$\widehat{E\max}_{tk} = g_t(\Omega_{tk}^-; \gamma_t) + \varsigma_{tk} \quad \text{for } t = 1, \ldots, t_m - 1. \tag{67}$$

In the binary case with additive normal errors, the cut-off values for the participation decision, which were the ingredients for the likelihood function calculation, were analytical. Moreover, although the likelihood function (35) did not have a closed form representation, it required the calculation only of a univariate cumulative normal distribution. In the multinomial choice setting we have described, the set of values of the ω vector determining optimal choices and serving as limits of integration in the probabilities associated with the work alternatives that comprise the likelihood function have no analytical form and the likelihood function requires a multivariate integration.

To accommodate these complications, maximum likelihood estimation of the model uses simulation methods. To describe the procedure, let the set of values of $\widetilde{\omega}_t$ for which the hnth choice is optimal at t be denoted by $S_t^{hn}(\Omega_{it}^-) = \{\omega_{1t}^1, \omega_{1t}^2, \omega_{1t}^3, \omega_{1t}^N, \omega_{2t}^w, \omega_{2t}^y | V_t^{hn} = \max(\widetilde{V_t^{hn}})\}$. Consider the probability that a couple chooses neither to work nor have a child, $h_{it} = 0, n_{it} = 0$, in a fecund period $t < t_m$:

$$\begin{aligned}&\Pr(h_{it} = 0, n_{it} = 0 | \Omega_{it}^-) \\ &\quad = \int_{S_t^{00}(\Omega_{it}^-)} f(\omega_{1t}^1, \omega_{1t}^2, \omega_{1t}^3, \omega_{1t}^N, \omega_{2t}^w, \omega_{2t}^y) \mathrm{d}\omega_{1t}^1 \mathrm{d}\omega_{1t}^2 \mathrm{d}\omega_{1t}^3 \mathrm{d}\omega_{1t}^N \mathrm{d}\omega_{2t}^w \mathrm{d}\omega_{2t}^y. \end{aligned} \tag{68}$$

This integral can be simulated by randomly taking $m = 1, \ldots, M$ draws from the joint distribution of ω, with draws denoted by ω_{mt}, and determining the fraction of times that the value function for that alternative is the largest among all eight feasible alternatives, that is,

$$\widehat{\Pr}(h_{it} = 0, n_{it} = 0 | \Omega_{it}^-) = \frac{1}{M} \sum_{m=1}^{M} 1[V_{it}^{00}(\widetilde{\omega_{mt}}) = \max(\widetilde{V_{mt}^{hn}}(\Omega_{it}^-))]. \tag{69}$$

One can similarly form an estimate of the probability for other nonwork alternatives, namely for $h_{it} = 0, n_{it} = 1$ for any $t < t_m$ and for $h_{it} = 0$ for any $t_m \leq t \leq T$. Recall that for infecund periods, there are only four alternatives because n_{it} is constrained to be zero.

When the wife works, the relevant probability contains the chosen joint alternative $\{h, n\}$ and the observed wage. For concreteness, consider the case where $h_{it} = 2$, $n_{it} = 1$. Then the likelihood contribution for an individual who works 2000 hours

[63] Interpolating functions should be chosen with great care. To avoid overfitting, it is useful to solve the model at more state points than used in estimating the interpolating function and use the additional points for cross-validation. For example, we might solve the model at 4000 state points, estimate the interpolating function on 2000 points and check the fit, say the R^2, using the other 2000 points.

in period t at a wage of w_{it} is

$$\Pr(h_{it}=2, n_{it}=1, w_{it}|\Omega_{it}^{-}) = \Pr(h_{it}=2, n_{it}=1|w_{it}, \Omega_{it}^{-}) \Pr(w_{it}|\Omega_{it}^{-}) \tag{70}$$

$$= \Pr(w_{it}|\Omega_{it}^{-}) \int_{S_t^{21}(\Omega_{it}^{-})} \mathrm{d}F(\omega_{1t}^1, \omega_{1t}^2, \omega_{1t}^3, \omega_{1t}^N, \omega_{2t}^y|\omega_{2t}^w). \tag{71}$$

For illustrative purposes, suppose that the (log) wage equation is additive in η_{it}^w,

$$\log w_{it} = \sum_{j=1}^{3} \log r^j d_{it}^j + \log \Psi_{it}(S_i^w, E_{it}, d_{it-1}^1, d_{it-1}^2, d_{it-1}^3) + \eta_{it}^w,$$

$$= \sum_{j=1}^{3} \log r^j d_{it}^j + \log \Psi_{it}(S_i^w, E_{it}, d_{it-1}^1, d_{it-1}^2, d_{it-1}^3)$$

$$+ \sum_{m^w=1}^{M} \lambda_{2m^w} 1(\text{type}^w = m^w) + \omega_{2it}^w \tag{72}$$

and further that $\widetilde{\omega}$ is joint normal.[64] With these assumptions, and denoting the deterministic part of the right hand side of (72) by $\overline{\log w_{it}}$, we can write

$$\Pr(h_{it}=2, n_{it}=1|w_{it}, \Omega_{it}^{-}) \Pr(w_{it}|\Omega_{it}^{-})$$

$$= \int_{S_t^{21}(\Omega_{it}^{-})} \mathrm{d}F(\omega_{1t}^1, \omega_{1t}^2, \omega_{1t}^3, \omega_{1t}^N, \omega_{2t}^y|\omega_{2t}^w = \log w_{it} - \overline{\log w_{it}})$$

$$\times \frac{1}{w_{it}\sigma_{\omega_2^w}} \phi\left(\frac{\log w_{it} - \overline{\log w_{it}}}{\sigma_{\omega_2^w}}\right) \tag{73}$$

where $\frac{1}{w_{it}}$ is the Jacobian of the transformation from the distribution of w to the distribution of ω_2^w. Under these assumptions $f(\omega_{1t}^1, \omega_{1t}^2, \omega_{1t}^3, \omega_{1t}^N, \omega_{2t}^y|\omega_{2t}^w)$ is normal and the frequency simulator for the conditional probability takes the same form as (69) except that ω_{2t}^w is set equal to $\log w_{it} - \sum_{j=1}^{3} \log r^j d_{it}^j + \log \Psi_{it} + \sum_{m^w=1}^{M} \lambda_{2m^w} 1(\text{type}^w = m^w)$ and the other five ω's are drawn from $f(\omega_1^1, \omega_1^2, \omega_1^3, \omega_1^N, \omega_2^y|\omega_2^w)$. Thus, denoting the fixed value of ω_{2t}^w as $\widehat{\omega_{2t}^w}$,

$$\Pr(h_{it}=2, n_{it}=1|w_{it}, \Omega_{it}^{-}) = \frac{1}{M} \sum_{m=1}^{M} 1\Big[V_{it}^{21}(\omega_{m1}^1, \omega_{m1}^2, \omega_{m1}^3, \omega_{m1}^N, \omega_{m2}^y, \widehat{\omega_{mt}^w})$$

$$= \max(\widetilde{V_{it}^{hn}}(\omega_{m1}^1, \omega_{m1}^2, \omega_{m1}^3, \omega_{m1}^N, \omega_{m2}^y, \widehat{\omega_{mt}^w}))\Big]. \tag{74}$$

[64] Note that the type-specific parameters, λ's, are essentially the constant terms in the Ψ production function and cannot be separately identified from the skill rental prices.

Although these frequency simulators converge to the true probabilities as $M \to \infty$, there is a practical problem in implementing this approach. Even for large M, the likelihood is not smooth in the parameters, which precludes the use of derivative methods (e.g., BHHH). This lack of smoothness forces the use of non-derivative methods, which converge more slowly. However, frequency simulators can be smoothed, which makes the likelihood function differentiable and improves the performance of optimization routines. One example is the smoothed logit simulator (McFadden, 1989), namely (in the case we just considered),

$$\Pr(h_{it} = 2, n_{it} = 1 | w_{i\tau}, \Omega_{it}^{-}) = \frac{1}{M} \sum_{m=1}^{M} \frac{\exp\left[(V_{itm}^{21} - \max(\widetilde{V_{itm}^{hn}}))/\tau\right]}{\sum_{\{h,n\}} \exp\left[(V_{itm}^{hn} - \max(\widetilde{V_{itm}^{hn}}))/\tau\right]} \quad (75)$$

where V_{itm}^{hn} is shorthand for the value functions in (74) and τ is a smoothing parameter. As $\tau \to 0$, the RHS converges to the frequency simulator. The other choice probabilities associated with work alternatives are similarly calculated.

3.2.1. Alternative estimation approaches

Conceptually, any dynamic programming problem that admits to numerical solution can be estimated. In addition to simulated maximum likelihood, researchers have used various alternative simulation estimation methods, including minimum distance estimation, simulated method of moments and indirect inference. There is nothing in the application of these estimation methods to DCDP models that is special, other than having to iterate between solving the dynamic programming problem and minimizing a statistical objective function.

The main limiting factor in estimating DCDP models is the computational burden associated with the iterative process. It is therefore not surprising that there have been continuing efforts to reduce the computational burden of estimating DCDP models. We briefly review two such methods.

A Bayesian approach

As has been discussed elsewhere (see Geweke and Keane (2000)), it is difficult to apply the Bayesian approach to inference in DCDP models because the posterior distribution of the model parameters given the data is typically intractably complex. Recently, however, computationally practical Bayesian approaches that rely on Markov Chain Monte Carlo (MCMC) methods have been developed by Imai et al. (2009) and Norets (2009). We will discuss the Imai et al. (2009) approach in the stationary case, where it is most effective. Thus, we remove time superscripts from the value functions and denote Ω' as the next period state. We also make the parameter vector θ explicit. Thus, corresponding to

Eq. (20) and the (21), we have

$$V(\Omega_i, \theta) = \max_d (V^0(\Omega_i, \theta), V^1(\Omega_i, \theta)), \tag{76}$$

where

$$V^k(\Omega_i, \theta) = U_i^k(\Omega_i, \theta) + \delta E[V(\Omega_i', \theta)|\Omega_i, d_i = k], \ldots k = 0, 1. \tag{77}$$

The basic idea is to treat not only the parameters but also the values functions and expected value functions as objects that are to be updated on each iteration of the MCMC algorithm. Hence, we add the superscript (s) to the value functions, the expected value functions and the parameters to denote the values of these objects on iteration (s). We use $\widehat{E^{(s)}}$ to denote the approximation to the expected value and $L(\theta^{(s)})$ to denote the likelihood.

The Imai et al. (2009) algorithm consists of three steps: the parameter update step (using the Metropolis-Hastings algorithm), the Dynamic Programming step, and the expected value approximation step:

(1) The Parameter Updating Step (Metropolis-Hastings algorithm)

First, draw a candidate parameter vector from the proposal density $\theta^{(s)*} \sim q(\theta^{(s)*}|\theta^{(s)})$. Then, evaluate the likelihood conditional on $\theta^{(s)*}$ and conditional on $\theta^{(s)}$. Now, form the acceptance probability

$$P = \min \left\{ \frac{L(\theta^{(s)*}) q(\theta^{(s)}|\theta^{(s)*})}{L(\theta^{(s)}) q(\theta^{(s)*}|\theta^{(s)})}, 1 \right\}. \tag{78}$$

We then accept $\theta^{(s)*}$ with probability P, that is,

$$\theta^{(s+1)} = \begin{Bmatrix} \theta^{(s)*} & \text{with probability } P \\ \theta^{(s)} & \text{....with probability } 1 - p \end{Bmatrix}. \tag{79}$$

(2) The Dynamic Programming (or Bellman equation iteration) Step

The following Bellman equation step is nested within the parameter updating step:

$$V^{(s)}(\Omega_i, \theta^{(s)}) = \max_d (V^{0(s)}(\Omega_i, \theta^{(s)}), V^{1(s)}(\Omega_i, \theta^{(s)})), \tag{80}$$

$$V^{k(s)}(\Omega_i, \theta^{(s)}) = U_i^k(\Omega_i, \theta^{(s)}) + \delta \widehat{E^{(s)}}[V(\Omega_i', \theta^{(s)})|\Omega_i, d_i = k], \ldots k = 0, 1. \tag{81}$$

The difficulty here is in obtaining the expected value function approximation that appears on the right hand side of (81). We describe this next.

(3) Expected value approximation step.

The expected value function approximation is computed using information from earlier iterations of the MCMC algorithm. The problem is that, on iteration (s), they have not, in general, yet calculated the value functions at the specific parameter value $\theta^{(s)}$ that they have drawn on iteration (s). Intuitively, the idea is to approximate the expected value functions at $\theta^{(s)}$ by looking at value functions that were already calculated on earlier iterations of the MCMC algorithm, emphasizing parameter values that are in some sense "close" to $\theta^{(s)}$.

Specifically, the expected value function is approximated as

$$\widehat{E^{(s)}}[V(\Omega_i', \theta^{(s)})|\Omega_i, d_i = k] = \frac{1}{N^{(s)}} \sum_{j=1}^{N^{(s)}} V^{(j)}(\Omega_i', \theta^{(j)}) W(\theta^{(j)}, \theta^{(s)}), \tag{82}$$

where $\theta^{(j)}$ denotes a parameter value from an earlier iteration (j) of the MCMC algorithm and $V^{(j)}(\Omega_i', \theta^{(j)})$ is the value function at state point Ω_i' that was calculated on iteration (j).[65] Finally, $W(\theta^{(j)}, \theta^{(s)})$ is a weighting function that formalizes the notion of closeness between $\theta^{(s)}$ and $\theta^{(j)}$. Imai et al. (2009) use weighting function given by

$$W(\theta^{(j)}, \theta^{(s)}) = \frac{K_h(\theta^{(j)}, \theta)}{\sum_{m=1}^{N^{(s)}} K_h(\theta^{(m)}, \theta)}, \tag{83}$$

where K_h is a kernel with bandwidth h.

Under certain conditions, as the number of iterations grows large, the output of this algorithm generates convergence to the posterior distribution of the parameter vector, as well as convergence to the correct (state and parameter contingent) value functions. One condition is "forgetting." That is, the algorithm will typically be initialized using rather arbitrary initial value functions. Hence, the sum in (82) should be taken using a moving window of more recent iterations so early iterations are dropped. Another key point is that, as one iterates, more lagged values of $\theta^{(j)}$ become available, so more values that are "close" to the current $\theta^{(s)}$ will become available. Hence, the bandwidth in the kernel smoother in (83) should become narrower as one iterates. Note that satisfying both the "forgetting" and "narrowing" conditions simultaneously requires that the "moving window" mentioned earlier must expand as one iterates, but not too quickly. Norets (2009) and Imai et al. (2009) derive precise rates.

The Bayesian methods described here are in principle applicable to non-stationary models as well. This should be obvious given that a non-stationary model can always be

[65] Note that in writing (82) we are implicitly assuming that the state space evolves deterministically, conditional on the current state and current choice. Otherwise (82) would require a double sum, where the inner sum is over states that could potentially be reached from Ω_i given the choice $d_i = k$. Norets (2009) handles the stochastically evolving state space case. See also Ching et al. (2010).

represented as a stationary model with (enough) age specific variables included in the state space. However, this creates the usual curse of dimensionality, as the state space may expand substantially as a result. Unlike, say, the approximate solution algorithm proposed by Keane and Wolpin (1994), these Bayesian algorithms are not designed (or intended) to be methods for handling extremely large state space problems. Combining the two ideas is a useful avenue for future research.

It is worth noting that no DCDP work that we are aware of has ever reported a distribution of policy simulations that accounts for parameter uncertainty; and, it is also rarely done in nonstructural work.[66] The Bayesian approach provides a natural way to do this, and Imai et al. (2009) have produced code that generates such a distribution.

A non-full solution method

Hotz and Miller (1993) developed a method for the implementing DCDP models that does not involve solving the DP model, that is, calculating the $E\max_t$ functions. HM prove that, for additive errors, the $E\max_t$ functions can be written solely as functions of conditional choice probabilities and state variables for any joint distribution of additive shocks. Although the method does not require that errors be distributed extreme value, the computational advantage of the method is best exploited under that assumption.

Consider again the binary choice model.[67] From (38), one can see that if we have an estimate of the conditional choice probabilities at all state points, $E\max_T$ can also be calculated at all state points. Denoting the (estimate of the) conditional choice probability by $\widehat{\Pr}(d_{iT}=1|\Omega_{iT}^{-})$,

$$\widehat{E}\max_T = \rho\left\{\gamma + \frac{y_T + z\gamma_1 + \gamma_2 h_T - \pi n}{\rho} - \log(\widehat{\Pr}(d_{iT}=1|\Omega_{iT}^{-}))\right\}. \tag{84}$$

Consider now period $T-1$ and suppose we have an estimate of the conditional choice probabilities, $\widehat{\Pr}(d_{iT-1}=1|\Omega_{iT-1}^{-})$. Then,

$$E\max_{T-1} = \rho\left\{\gamma + \frac{y_{T-1} + z\gamma_1 + \gamma_2 h_{T-1} - \pi n + \delta\widehat{E}\max_T(h_{T-1}+1)}{\rho} - \log(\widehat{\Pr}(d_{iT-1}=1|\Omega_{iT-1}^{-}))\right\}, \tag{85}$$

where, for convenience, we have included only work experience in the $\widehat{E}\max_T$ function. We can continue substituting the estimated conditional choice probabilities in this recursive manner, yielding at any t

[66] Of course, providing such a distribution is possible without adopting a Bayesian approach, although it can be computationally burdensome.

[67] The main insight in the multinomial setting is the same and the extension is straightforward.

$$\widehat{E}\max_t = \rho\left\{\gamma + \frac{y_t + z\gamma_1 + \gamma_2 h_t - \pi n + \delta\widehat{E}\max_{t+1}(h_t+1)}{\rho} - \log(\widehat{\Pr}(d_{it}=1|\Omega_{it}^-))\right\}. \quad (86)$$

These $\widehat{E}\max_t$ functions can be used in determining the $\xi_{it}^*(\Omega_{it}^-)$ cut-off values that enter the likelihood function.

As with other approaches, there are limitations. First, the empirical strategy involves estimating the conditional choice probabilities from the data (nonparametrically if the data permit). In the case at hand, the conditional choice probabilities correspond to the proportion of women who work for given values of the state variables (for example, for all levels of work experience). To implement this procedure, one needs estimates of the conditional choice probabilities through the final decision period and for each possible value of the state space. Thus, we need longitudinal data that either extends to the end of the decision period or we need to assume that the conditional choice probabilities can be obtained from synthetic cohorts. This latter method requires an assumption of stationarity, that is, in forecasting the conditional choice probabilities of a 30 year old observed in year t when reaching age 60 in year $t+30$, it's assumed that the 30 year old would face the same decision-making environment (for example, the same wage offer function, etc.) as the 60 year old observed in year t. Most DCDP models in the literature which solve the full dynamic programming problem implicitly make such an assumption as well, though it is not dictated by the method.[68] Moreover, it must also be assumed that there are no state variables observed to the agent but unobserved to us; otherwise, we will not be matching the 30 year olds to the 60 year olds the same unobserved state values.[69] Second, the convenience of using additive extreme value errors brings with it the previously discussed limitations of that assumption. Third, the estimates are not efficient, because the fact that the $\widehat{E}\max_t$ functions themselves contain the parameters in the model structure are not taken into account.

4. APPLICATIONS

In this section we review the substantive contributions of the DCDP literature to three main areas of labor economics: (i) labor supply (female and male), (ii) job search and (iii) human capital.

68 Lee and Wolpin (2006); Lentz (2009) allow for (equilibrium) skill prices to change with calendar time due to technical change.

69 In recent work, Arcidiacono and Miller (2008) have developed methods for extending the HM approach to allow for unobserved state variables. However, there has as yet been no empirical implementation of that approach to a model as rich as those found in the literature.

4.1. Labor supply

The literature on dynamic labor supply models can be usefully divided into that on females and males. This is because the two literatures have emphasized different aspects of behavior. A key feature of female labor supply is that a large percentage of women (particularly married women) do not work during significant portions of their life cycle. The central role of the decision of whether or not to work has made the DCDP approach more common in the study of female labor supply than in the literature on males.

The literature on women has also emphasized the relationship between participation and human capital accumulation, while tending to ignore saving. This is no accident, because, as Eckstein and Wolpin (1989b) note, it is very difficult computationally to handle participation, human capital and saving simultaneously.[70] The literature has also striven to model how fertility, marriage and participation decisions interact.

In contrast, the literature on males has emphasized the continuous choice of hours of work and savings, with participation usually taken as given. Given an assumption of interior solutions, most papers on dynamics of male labor supply have worked with the first order conditions of agents' optimization problems, rather than using the DCDP approach.[71] Nevertheless, at the end of this section, we review an empirical paper on male labor supply (Imai and Keane, 2004) that adapts the DCDP approach to the case of continuous choices of labor supply and consumption.

4.1.1. Female labor supply

As we have already noted in the previous discussion, the prevalence of nonparticipation creates a problem for the analysis of labor supply decisions given that a person's market wage rate is not usually observed for nonparticipants. The classic paper by Heckman (1974) developed a method for estimating a labor supply function (with continuous hours and nonparticipation) when wages are only observed for workers. In his framework, the labor supply function is estimated jointly with a wage offer function by maximum likelihood.

The possibility of nonparticipation raises several additional issues. First, participating in the labor market may entail a fixed time and/or money cost (Cogan, 1981). Second, nonparticipation may lead to a lack of skill appreciation. Thus, the literature on female labor supply has allowed work experience, as a measure of human capital accumulated on

[70] Indeed, to our knowledge the only paper that has attempted to do so is Keane and Wolpin (2001). That paper models the labor supply and human capital investment decisions of young men, who often have low participation rates.

[71] For men, strict application of the DCDP approach would require discretization of hours as an approximation to the choice set. In that case, the parallel to the multinomial choice problem considered above is exact. However, the main insight of the DCDP approach to estimation applies as well to continuous choices and to discrete-continuous choices in which the underlying dynamic programming problem is solved based on first-order conditions or Kuhn-Tucker conditions. That insight was simply the observation that because the continuation value (the E max function) is a deterministic function of state variables, the static model and the dynamic programming model have a common empirical structure.

the job, to affect wage offers (e.g., Weiss and Gronau, 1981; Eckstein and Wolpin, 1989b). Third, the fact that there is heterogeneity in the extent to which women participate over their lifetimes raises the question of the extent to which that heterogeneity is due to permanent (unobserved) differences in preferences for work or to the influence of past work decisions on participation that arise through transitory taste shocks (Heckman and Willis, 1977). Fourth, nonparticipation implies a potentially central role for marriage and fertility decisions.

4.1.2. Mincer's (1962) life cycle model

The earliest paper on labor supply of women to adopt formally a life cycle perspective was Mincer (1962). A married woman's labor supply is based on the permanent income of her husband, as well as her market wage and the couple's tastes for market work, home work and children. Given this framework, the observed variation over the life-cycle in a woman's work hours is merely the result of the allocation of work hours to periods when market wages are high relative the value of home time (i.e., intertemporal substitution).

Based on this framework, Mincer (1962) hypothesized that a transitory change in husband's income, which has no significant effect on his permanent income, should have no impact on a woman's labor supply. Mincer provided some informal evidence on this hypothesis using data from the 1950 Survey of Consumer Expenditures. Taking 6,766 married white women, he stratified them into 12 groups based on husband's education and age and on the presence of young children. He then subdivided each group into households where the husband worked all year vs. those where the husband had a spell of unemployment. Mincer found that women had a higher participation rate if the husband had experienced an unemployment spell. Based on this evidence that women do respond to transitory changes in husband's income, Mincer concluded that a simple life-cycle model (with perfect foresight and no constraints on borrowing) could not adequately describe the data.

Two points are worth noting. First, Mincer (1962) uses households where the husband works all year as a "control group" for similar households (in terms of education, age and children) where the husband experiences an unemployment spell, with unemployment as the "treatment." Thus, one possible explanation for Mincer's finding is that the treatment and control groups differ in unobserved ways, and that women in the treatment group would have worked more regardless. Second, there are alternative explanations that are consistent with a life-cycle model. For instance, depending on the stochastic process for husband's income, unemployment shocks may induce long lived reductions in earnings. It is also possible that leisure time of the husband and wife are nonseparable in utility or that unemployed husbands may contribute to home production and/or child care. In either case, unemployment of the husband may reduce the value of home time for the wife.

4.1.3. Non-full solution methods of estimation

The modern structural literature on the estimation of life cycle models of female labor supply begins with Heckman and MaCurdy (1980).[72] They adopt the utility function

$$U_{it} = \alpha_{it} C_{it}^{\eta} + \beta_{it}(H_{\max} - h_{it})^{\gamma} \quad 0 < \eta < 1, 0 < \gamma < 1, \tag{87}$$

where C_{it} is household i's consumption at t, h_{it} the wife's hours of work, $H_{\max}$ maximum available hours in the period, and α_{it} and β_{it} are taste-shifters. Leisure is given by $L_{it} = (H_{\max} - h_{it})$. Households have perfect foresight about future preferences and wages. The household maximizes its discounted flow of utility over the finite horizon, $t = 0, \ldots, T$,

$$V_i = \sum_{t=0}^{T} \frac{1}{(1+\rho)^t} U_{it}, \tag{88}$$

where ρ is the household's subjective rate of time preference. The household faces the lifetime budget constraint

$$A_{i0} + \sum_{t=0}^{T} \frac{w_{it} h_{it}}{(1+r)^t} = \sum_{t=0}^{T} \frac{C_{it}}{(1+r)^t}, \tag{89}$$

where A_{i0} is the household's initial assets and r is the (constant) rate of interest. Assuming an interior solution, the first-order conditions for all $t = 0, \ldots, T$ are

$$\frac{\partial U_{it}}{\partial C_{it}} = \left[\frac{(1+\rho)}{(1+r)}\right]^t \lambda_i, \tag{90}$$

$$\frac{\partial U_{it}}{\partial L_{it}} = \left[\frac{(1+\rho)}{(1+r)}\right]^t w_{it} \lambda_i, \tag{91}$$

where λ_i is the marginal utility of wealth at $t = 0$. Using the utility function specification (87), (91) becomes

$$\gamma \beta_{it} L_{it}^{\gamma - 1} = \left[\frac{(1+\rho)}{(1+r)}\right]^t w_{it} \lambda_i. \tag{92}$$

Taking logs and rearranging yields the Frisch demand function for leisure,

$$\log L_{it} = \frac{1}{\gamma - 1} \left\{ \log w_{it} + \log \lambda_i + t \log \left[\frac{(1+\rho)}{(1+r)}\right] - \log \beta_{it} - \log \gamma \right\}. \tag{93}$$

[72] Their approach builds on the seminal work of MaCurdy (1981) on the labor supply of men.

To deal with corner solutions, Heckman and MaCurdy (1980) note that a women will choose not to work if the marginal utility of leisure, evaluated at zero hours of work, exceeds the marginal value of working, that is, if

$$\left.\frac{\partial U_{it}}{\partial L_{it}}\right|_{L_{it}=H_{\max}} > \left[\frac{(1+\rho)}{(1+r)}\right]^t \lambda_i w_{it}, \text{ or} \tag{94}$$

$$\gamma \beta_{it} H_{\max}^{\gamma-1} > \left[\frac{(1+\rho)}{(1+r)}\right]^t \lambda_i w_{it}. \tag{95}$$

Taking logs and rearranging, we can express this participation condition as a reservation wage condition, namely

$$h_{it} > 0 \quad \text{iff}$$
$$\log w_{it} > -\log \lambda_{i0} - t \log\left[\frac{(1+\rho)}{(1+r)}\right] + \log \beta_{it} + \log \gamma - (1-\gamma)\log H_{\max}. \tag{96}$$

Notice that if the household has a lower level of lifetime wealth, and hence a higher value of λ_i, the reservation wage is correspondingly reduced.

To obtain an estimable model, Heckman and MaCurdy (1980) assume functional forms for the taste shifter β_{it} and for the wage offer function, namely

$$\log \beta_{it} = X_{it}\phi + \eta_{1i} + \epsilon_{1it}, \tag{97}$$
$$\log w_{it} = Z_{it}\theta + \eta_{2i} + \epsilon_{2it}, \tag{98}$$

where X_{it} and Z_{it} are vectors of observables that affect the taste for leisure and market productivity, η_{1i} and η_{2i} are individual permanent components of the taste for leisure and market productivity and ϵ_{1it} and ϵ_{2it} are respective transitory shocks. Substituting (97) and (98) into (93) and (96), we obtain reduced form equations for (i) leisure conditional on participation and (ii) the participation decision rule:

$$\log L_{it} = f_i + X_{it}\frac{\phi}{\gamma-1} - Z_{it}\frac{\theta}{\gamma-1} + \frac{1}{\gamma-1}\log\left[\frac{(1+\rho)}{(1+r)}\right] t + \frac{\epsilon_{2it}-\epsilon_{1it}}{\gamma-1}, \tag{99}$$

$$h_{it} > 0 \quad \text{iff} \quad \frac{\epsilon_{2it}-\epsilon_{1it}}{\gamma-1} < -f_i - X_{it}\frac{\phi}{\gamma-1} + Z_{it}\frac{\theta}{\gamma-1}$$
$$-\frac{1}{\gamma-1}\log\left[\frac{(1+\rho)}{(1+r)}\right] t + \log H_{\max}, \tag{100}$$

where $f_i = \frac{1}{\gamma-1}\{\log \lambda_i + \eta_{2i} - \eta_{1i} - \log \gamma\}$ is an individual-specific fixed effect which subsumes the marginal utility of wealth term λ_i as well as the individual permanent components of tastes for work and productivity.

Under the assumptions of the model (i.e., perfect foresight, no borrowing constraints) this fixed effect is time invariant, capturing everything from periods outside of period t

relevant for the woman's labor supply decision at time t. For example, in this model it is not necessary to explicitly include the current or future earnings of a married woman's husband, which is captured through λ_i. In principle, it is not even necessary to control explicitly for whether a woman is married, as the woman's marriage history is also built into λ_i. For instance, a single woman is assumed to anticipate the earnings potential of any husband she will eventually marry. Marriage can only enter the model because it shifts tastes for work, not because it alters lifetime wealth.

To estimate the model Heckman and MaCurdy (1980) assume that the stochastic terms ϵ_{1it} and ϵ_{2it} are jointly normal and serially uncorrelated.[73] The hours and participation Eqs (99) and (100) are estimated jointly with the wage Eq. (97) by maximum likelihood. The data consist of 30 to 65 year old continuously married white women from the 1968-75 waves of the PSID. There are 672 women who meet the selection criteria, but to estimate the fixed effects f_i, only women who work at least once can be used, leaving 452.[74]

The variables included in the wage equation Z_{it} are potential experience (i.e., age—education—6) and its square along with the local unemployment rate. Because only time varying covariates can be included due to the presence of the fixed effect in the wage equation, education, for example, is not included. The variables included as taste shifters X_{it} are the total number of children, the number of children less than 6, the wife's age, a measure of the number of hours the husband is unemployed, "other" household income, and an indicator for whether the husband is retired or disabled.[75]

The results of the estimation are mostly standard. Tastes for home time are increasing in the number of children and especially the number less than 6. Both "other" income and the husband's hours of unemployment are statistically insignificant, which Heckman and MaCurdy (1980) interpret as evidence that supports the life-cycle model and that contradicts Mincer (1962). But interestingly, the estimate of γ bumps up against its lower bound of zero. This implies a Frisch elasticity of leisure of $\frac{1}{\gamma-1} = -1$. Converting to a Frisch labor supply elasticity, and noting that mean hours worked in the sample is about 1300, we have that

$$\begin{aligned}\frac{\partial \log h_{it}}{\partial \log w_{it}} &= \frac{\partial \log h_{it}}{\partial \log L_{it}} \frac{\partial \log L_{it}}{\partial \log w_{it}} = \frac{L_{it}}{H_{\max} - L_{it}} \frac{1}{1-\gamma} \\ &\approx \frac{L_{it}}{h_{it}} = \frac{7460}{1300} = 5.7 \end{aligned} \tag{101}$$

which is certainly a large value.

[73] $H_{\max}$ is set at 8760 hours.

[74] If a woman never works, the likelihood of that event is maximized by setting the fixed effect to $-\infty$. Adjustments for this sample selection made little difference to the estimates.

[75] Although the effect of the wife's age may be interpreted as an estimate of $\log\left[\frac{(1+\rho)}{(1+r)}\right]$, it may also reflect changing preferences for leisure with age.

In a subsequent paper, Heckman and MaCurdy (1982) acknowledged that their choice of functional form had implicitly constrained the elasticity of substitution for leisure, and also for hours, to be large. Specifically, if the Frisch elasticity for leisure is $\frac{1}{\gamma-1}$ and we impose $0 < \gamma < 1$, then the elasticity must range from -1 to $-\infty$. Then, for example, if leisure takes up at least two thirds of available time, (101) implies that the Frisch elasticity of labor supply must be at least 2.

Heckman and MaCurdy (1982) report new results based on an additively separable CRRA utility function,

$$U_{it} = \alpha_{it}\eta^{-1}C_{it}^{\eta} + \beta_{it}\gamma^{-1}(H_{\max} - h_{it})^{\gamma} \quad \eta < 1, \gamma < 1. \tag{102}$$

Adopting (102) does not change anything important in terms of the estimating equations, the only difference being that the constant term $\log\gamma$ drops out of the equation for f_i. But now the constraint on γ is only that it be less than one. In fact, Heckman and MaCurdy (1982) estimate $\gamma = -1.44$, which implies a Frisch elasticity of leisure of $\frac{1}{\gamma-1} = -.41$. Interestingly, this still implies a large value of the Frisch elasticity of labor supply equal to 2.35.

The change in the utility specification has some impact on the other parameter estimates. The impact of children on tastes for work becomes larger. The coefficient on income of other household members becomes quantitatively much larger, but is only significant at the 20% level. Heckman and MaCurdy (1982) interpret this result as being "less favorable toward the permanent income hypothesis." Husband unemployment hours also becomes marginally significant and negative, implying that husband time at home increases the wife's tastes for work.

Heckman and MaCurdy (1980) conduct a second stage estimation where they regress the fixed effects on various determinants of lifetime wealth. Given estimates of the fixed effects, f_i, and given an estimate of γ and the wage equation fixed effects η_{2i}, we can back out estimates of $\log\lambda_i - \eta_{1i}$. Thus, it is possible to isolate only a composite of the marginal utility of wealth minus the fixed effect in tastes for leisure. It turns out that this composite is reduced by wife's education. We would expect education to increase lifetime wealth (thus reducing λ_i) both by increasing own and potential husband's earnings. But the effect of education on tastes for leisure η_{1i} is an empirical question. The result implies either that education increases taste for leisure, or, if it reduces it, that this effect is outweighed by the income effect.

The Heckman and MaCurdy (1980, 1982) papers, as well as earlier work in a static framework by Heckman (1974), do not accommodate fixed costs of work. Within a static model, Cogan (1981) argued that ignoring fixed costs can lead to severe bias in estimates of female labor supply functions. To see the problem, consider the simple quasi-linear utility function given by

$$U = C + \beta \frac{(H_{\max} - h)^{1+\gamma}}{1+\gamma}$$
$$= (wh + Y - F) + \beta \frac{(H_{\max} - h)^{1+\gamma}}{1+\gamma}, \tag{103}$$

where Y represents non-labor income and F represents fixed costs of working (e.g., child care costs). The equation for optimal hours conditional on working is simply

$$h^* = H_{\max} - \left(\frac{w}{\beta}\right)^{1/\gamma}. \tag{104}$$

In the absence of fixed costs the reservation wage (w_R) is

$$h^* > 0 \Rightarrow H_{\max} - \left(\frac{w}{\beta}\right)^{1/\gamma} > 0 \quad \text{or} \tag{105}$$
$$w > \beta H_{\max}^{\gamma} = w_R.$$

However, as Cogan (1981) points out, it is not appropriate to use marginal conditions to determine the participation decision rule in the presence of fixed costs. Instead, it is necessary to compare the utilities conditional on working and not working, that is,

$$U(h^*) = w\left[H_{\max} - \left(\frac{w}{\beta}\right)^{1/\gamma}\right] + Y - F + \frac{\beta}{1+\gamma}\left[\left(\frac{w}{\beta}\right)^{1/\gamma}\right]^{1+\gamma} \quad \text{and} \tag{106}$$
$$U(0) = Y + \frac{\beta}{1+\gamma} H_{\max}^{1+\gamma}.$$

Thus, the decision rule for whether to work (whether $U(h^*) > U(0)$) can be expressed as

$$h > 0 \quad \text{iff} \quad h^* = \left[H_{\max} - \left(\frac{w}{\beta}\right)^{1/\gamma}\right]$$
$$> \frac{F}{w} + \frac{1}{w}\frac{\beta}{1+\gamma}\left\{H_{\max}^{1+\gamma} - \left[\left(\frac{w}{\beta}\right)^{1/\gamma}\right]\right\}^{1+\gamma}$$
$$= h_R > 0. \tag{107}$$

It is instructive to compare (105), which simply says that the person begins to work when desired hours are positive with (107), which says a person will begin to work only when optimal hours cross a positive threshold value h_R, which Cogan (1981) refers to as reservation hours. Inspection of the right hand side of the inequality in (107) provides

intuition for the threshold value; optimal hours conditional on working must be high enough to cover fixed costs plus an additional term which equals the monetized value of the lost utility from leisure.

Thus, as Cogan (1981) describes, in the presence of fixed costs of work the labor supply function is discontinuous, jumping from zero to the reservation hours level when the reservation wage is reached. The specifications assumed in Heckman (1974) and Heckman and MaCurdy (1980, 1982) are not consistent with such behavior. Another key point is that both costs of working (F) and tastes for work (β) enter the participation equation, while only β enters the labor supply equation. Hence, it is possible that a variable like young children could affect fixed costs of working but not tastes for work, that is, that the presence of young children could affect the participation decision but not labor supply conditional on participating.

To estimate labor supply behavior in the presence of fixed costs, Cogan (1981) proposes to jointly estimate a labor supply function as in (104), a reservation hours function as in (107) and an offer wage function. This is in contrast to Heckman's approach of jointly estimating a labor supply function (104), a participation equation based on marginal conditions as in (105) and an offer wage function.

Cogan (1981) compares both approaches using data on married women aged 30 to 34 taken from the 1967 National Longitudinal Survey of Mature Women. In this sample, 898 wives worked and 939 did not. The labor supply and reservation hours functions both include the wife's education and age, number of young children, and husband's earnings. Cogan estimates that fixed costs are substantial (about 28% of average annual earnings), and that a young child raises fixed costs by about a third. He finds that ignoring fixed costs leads to severe overestimates of labor supply elasticities (conditional on work). Cogan's labor supply function implies a Marshallian elasticity of 0.89 at the mean of the data, compared to 2.45 obtained using the Heckman (1974) approach. The Hicks elasticities are 0.93 vs. 2.64.

However, Cogan also shows that the elasticities are rather meaningless in this context. As he notes, a 10% increase in the offer wage to the average nonworking woman in the sample would not induce her to enter the labor market. But a 15% increase would induce her to jump to over 1300 hours. However, an additional 15% wage increase would "only" induce a further increase of 180 hours (or 13.6%).[76]

An important aspect of Cogan (1981) is that he pays close attention to how the model fits the distribution of hours. This is quite unusual in the static literature, where the focus tends to be on estimating elasticities rather than simulating behavior.[77] Cogan finds

[76] Note that this is still a rather large increase, consistent with a Marshallian elasticity of $13.6/15 = 0.90$.

[77] The only exceptions we have come across are Van Soest et al. (1990) and Keane and Moffitt (1998). Both papers note that it is rare to observe people working very low levels of hours (the former paper looking at men, the latter looking at single mothers). Van Soest et al. (1990) capture this by building in a job offer distribution where few jobs with low levels of hours are available. Keane and Moffitt (1998) build in actual measures of fixed costs of working (e.g., estimates of child care costs).

that the model without fixed costs cannot explain the fact that few people are observed to work very few hours. Indeed, the model without fixed costs has to predict a large fraction of women working few hours to be able to fit the large fraction of women who do not work. As Cogan describes, this leads to a flattening of the labor supply function, which exaggerates wage elasticities (see Cogan, 1981, Fig. 2). The model with fixed costs provides a much better fit to the data and does not have this problem.

Kimmel and Knieser (1998) extend the Heckman and MaCurdy (1980, 1982) analysis to include fixed costs of work. That is, they estimate a labor supply equation analogous to (99) jointly with a participation decision rule and an offer wage function, namely

$$\log h_{it} = f_{hi} + e_F \log w_{it} + \alpha_h X_{it} + \epsilon_{hit}, \tag{108}$$

$$\Pr(h_{it} > 0) = \Phi(f_{pi} + \beta \log w_{it} + \alpha_p X_{it}). \tag{109}$$

The first equation is the Frisch labor supply function where the fixed effect f_{hi} captures the marginal utility of initial assets along with any fixed effects in tastes for work. The second equation gives the probability of participation, where Φ is the cumulative standard normal. The fixed effect f_{pi} captures not just the marginal utility of wealth and tastes for work, but also individual heterogeneity in the fixed costs of work.

Following Cogan (1981), the existence of fixed costs breaks the tight link between the parameters in the participation and labor supply equations as we previously saw. Thus, there is no necessary relationship between the parameters e_F and α_h in (108) and the parameters β and α_p in (109). In this framework e_F is the conventional Frisch elasticity of labor supply conditional on employment. But, we can also introduce a Frisch participation elasticity given by

$$e_P = \frac{\partial \log \Pr(h_{it} > 0)}{\partial \log w_{it}} = \beta \frac{\phi(\cdot)}{\Phi(\cdot)}, \tag{110}$$

where ϕ is the standard normal density.

Kimmel and Knieser (1998) estimate this model using data on 2428 women from the Survey of Income Program Participation (SIPP), 68% of whom are married. The tri-annual interview information was collected in May 1983 to April 1986, giving 9 periods of data. The variables included in X_{it} are marital status, children, education and a quadratic in time. The model is estimated in two stages, where in the first stage predicted wages are constructed for workers and nonworkers by estimating the wage equation using Heckman's (1979) two-step procedure. The use of predicted wages serves three purposes: (i) to deal with measurement error, (ii) to fill in missing wages and (iii) to deal with possible endogeneity of wages (which would arise if women with high unobserved tastes for work also tend to have high wages). The variables that appear in the wage equation but not in X_{it} are race and a quadratic in age (potential experience).

The estimates imply a Frisch elasticity of 0.66 for employed women, and a Frisch participation elasticity of 2.39. Average hours of the entire population is given by $\overline{h} = P\overline{h}_e$, where $\overline{h}_e$ is average hours of the employed and P is the percentage employed. Thus we have that

$$\frac{\partial \log \overline{h}}{\partial \log w} = \frac{\partial \log P}{\partial \log w} + \frac{\partial \log \overline{h_e}}{\partial \log w} = 0.66 + 2.39 = 3.05.$$

Thus, the participation elasticity is much larger than the hours elasticity. This result provides some justification for models of female labor supply that focus primarily on the participation decision (see below).

Altug and Miller (1998) extend the life-cycle model of Heckman and MaCurdy (1980, 1982) to include human capital accumulation in the form of learning-by-doing. In addition, they incorporate fixed costs of work, state dependence in tastes for leisure, and aggregate shocks. The first step in Altug and Miller (1998) is to estimate the wage offer function, which takes the form

$$\log w_{it} = \log \omega_t + Z_{it}\gamma + \nu_i + \eta_{it}. \tag{111}$$

Here Z_{it} is a vector containing work experience, lagged participation and hours, and other observable determinants of skill, ν_i is a time-invariant skill endowment of person i and ω_t is a skill rental price (determined in equilibrium). In estimation, the ν_i can be treated as individual fixed effects and $\log \omega_t$ as time dummies. A key assumption is that η_{it} reflects only measurement error (and not unobserved variation in skill). Given that assumption, no selection bias problem arises if we estimate (111) by OLS only using periods when women are working, provided we include fixed effects.

Altug and Miller (1998) estimate the wage offer function using PSID data from 1967 to 1985. They require that the women reside in a PSID household for at least 6 consecutive years and that they be employed for at least two years (so that the fixed effects, ν_i, can be estimated). This gives a sample of 2169 women. The estimates imply that labor market experience, particularly recent experience, has a large effect on current wages. For instance, a person who worked the average level of hours for the past four years would have current offer wages about 25% higher than someone who had not worked. Interestingly, the lagged participation coefficients are negative while lagged hours coefficients are positive. The implication is that low levels of hours do not increase human capital: one has to work about 500 to 1000 hours to keep skill from depreciating.

The time dummies in the estimation are estimates of the rental price of skill. The rental price is estimated to be pro-cyclical, falling in the recession years of 1975 and 1980-1982 and rising in 1977, 1983 and 1985. Average wages among all women in the PSID sample are slightly more pro-cyclical than the estimated rental rates. This suggests

a compositional effect whereby people with high ν_i's tend to enter during booms. This is consistent with the mild pro-cyclical bias in aggregate wage measures for males found by Keane et al. (1988).

Altug and Miller (1998) assume a current period utility function given by

$$U_{it} = \alpha_{it}\eta^{-1}C_{it}^{\eta} + d_{it}[U_0(X_{0it}) + U_1(X_{1it}, h_{it}) + \epsilon_{1it}] + (1 - d_{it})\epsilon_{0it}. \quad (112)$$

Here the first term is CRRA in consumption, d_{it} is an indicator for positive hours, $U_0(\cdot)$ captures the fixed cost of work, $U_1(\cdot)$ is the disutility of labor, X_{oit} is a vector of demographic variables that reflect the fixed costs of working, X_{1it} includes X_{0it} along with lagged hours of work that shift tastes for leisure hours and ϵ_{1it} and ϵ_{0it} are stochastic shocks to tastes for the work and nonwork options, respectively. These shocks can be interpreted as unobserved variation in the fixed cost of work and the value of home time. Additive separability and the distributional assumptions on ϵ_{1it} and ϵ_{0it} play a key role in the estimation procedure, as discussed below.

As in an earlier paper (Altug and Miller, 1990), it is assumed that markets are complete (that all idiosyncratic shocks are perfectly insurable). Given this assumption and the specification of the utility function, the marginal utility of consumption can be shown to be given by

$$\alpha_{it}C_{it}^{\eta-1} = \lambda_{it} = \zeta_i\lambda_t \quad \text{or} \quad (113)$$

$$\log C_{it} = \frac{1}{\eta - 1}(\log\zeta_i + \log\lambda_t - \log\alpha_{it}). \quad (114)$$

As seen in (113), perfect insurance implies that λ_{it} can be decomposed into the product of an individual-specific component ζ_i, reflecting the marginal utility of wealth for individual i, and a time varying component λ_t, reflecting aggregate shocks. A person i with a low ζ_i has a relatively low marginal utility of wealth. But, a person's position in the wealth distribution is constant over time. The only source of uncertainty in the marginal utility of wealth over time are aggregate shocks that cause movements in λ_t.

To obtain an estimable equation, let $\log\alpha_{it} = X_{it}\beta + \epsilon_{cit}$, where X_{it} and ϵ_{cit} are observed and unobserved shifters of tastes for consumption, respectively. The consumption equation, (114), can be estimated by fixed effects (or in first differences), assuming the X_{it} are exogenous. Altug and Miller (1998) include household size, children, age and region in X_{it} and the λ_t are estimated as time dummies. The equation is estimated on data from the PSID, which contains only food consumption. As we would expect, the estimated values of λ_t are high in the recession years of 1975 and 1980–1982.

In the final step, Altug and Miller (1998) estimate the first order condition for hours jointly with a participation condition which allows for fixed costs of work. The first order condition for hours is complex because the marginal utility of leisure is not equated to simply the current wage times the marginal utility of consumption. There is an additional

term that arises because working today increases future wages and alters future disutilities from work. We refer to this term as the "expected future return to experience."

Altug and Miller deal with this problem using a version of the Hotz and Miller (1993) estimation algorithm.[78] To outline that procedure, first, given estimates of (111) and (114), they back out estimates of the individual effects ν_i and ζ_i. Second, they use nonparametric regression to estimate the probabilities of participation conditional on the state variables, that is, on the estimated values of ν_i and ζ_i, the work history, and a set of demographics (age, education, marital status, race, children, age and region).[79] Third, they assume the ϵ_{1it} and ϵ_{0it} in (112) are iid extreme value shocks, noting that they are the only source of randomness in the current period payoffs from working vs. not working. As in Hotz and Miller (1993), the value functions at any state can be backed out from the conditional choice probabilities calculated in step 2. This allows one to express the "expected future return to experience" terms as a simple function of the conditional participation probabilities (and their derivatives with respect to h_{it}). In the final estimation step, the parameters left to be estimated are the those associated with the fixed cost of work $U_0(X_{0it})$ and the disutility of labor $U_1(X_{1it}, h_{it})$.

It is important to understand the restrictions in this approach. There can be no stochastic variation in the marginal utility of leisure, because this additional source of randomness would preclude obtaining simple expressions for the expected future return to experience. Having actual productivity shocks instead of only measurement error in wages would have the same effect. And, consumption and leisure must be separable in utility, so that the stochastic term in tastes for consumption does not influence labor supply decisions. Thus, the extreme value error and additive separability assumptions are crucial.

So far, we have discussed approaches based on estimating the first-order condition for optimal labor supply. An alternative is the "life-cycle consistent" or "two-stage budgeting" approach, where one estimates labor supply equations that condition on the full income allocated to a period (MaCurdy, 1983). Using this approach, Blundell and Walker (1986) estimate a life-cycle consistent model of labor supply behavior of married couples. They use data on couples where both the husband and wife work, and the estimation of the labor supply function is done jointly with a probit equation for whether the wife works (to control for selection into the sample). In sharp contrast to Heckman and MaCurdy (1982) and Kimmel and Knieser (1998), they obtained an (average) Frisch elasticity of labor supply for women of only 0.033. The Hicks elasticity is 0.009. Based on the figures in their paper, we calculate an income effect of −0.206 (at the mean of the data) and a Marshallian elasticity of −0.197.

78 See Altug and Miller (1998), Eqs. (6.8) and (6.9), which give the final simple expressions for the labor supply and participation equations. Hotz et al. (1994) develop a simulation method for implementing the Hotz and Miller (1993) conditional choice probability approach.

79 Note it is important not to include the aggregate prices λ_t and ω_t in these regressions. Agents are assumed not to know the future realizations of these variables and so cannot condition on them when forming expected future payoffs.

Blundell et al. (1998) applied this life-cycle consistent approach to married women from UK Family Expenditure Survey 1978 to 1992. UK tax rates were reduced substantially over the period, and the basic idea of the paper is to exploit this variation to help identify labor supply elasticities. As the authors describe, the decline in rates caused different cohorts to face different paths of tax rates. Relative wages for different education groups also changed markedly over this period.

The idea of the paper can be understood as follows. Imagine we group the data by cohort and education level. That is, for each education/cohort we construct group means of hours and wages in each year. We then subtract group and time means from these quantities. The key assumption in Blundell et al. (1998) is that any residual variation in wages after taking out group and time means is exogenous. Their leading example of what might cause such residual variation in wages for a group is tax changes that affect groups differentially. Another source of variation would be exogenous technical change that affects groups differently. The key assumption here is that there are no shifts in labor supply behavior within any of the groups over time (e.g., tastes for leisure can vary by cohort/education level, but not within an education/cohort group over time). They also assume that taking out time means purges both hours and wages for all groups from the influence of aggregate shocks, a seemingly strong assumption as time affects (like the business cycle) may well affect different education/skill groups differently.

The simplest way to think about using the grouped data is to think of regressing the group mean of hours on the group mean of wages, after purging these means of group and time effects. An equivalent approach is to use the individual data and proceed in two steps. In the first step regress after-tax wages on time/group interaction dummies, and get the residuals from this regression. In the second step, regress hours on the after-tax wage, time and group dummies and wage residual. Note that we want the wage coefficient to be identified by wage variation within group over time. The wage equation residual captures other sources of wage variation, as the first stage wage equation controlled for time/group interactions.[80]

The authors also attempt to deal with possible compositional effects of changes in participation rates on the mean of the error term in the labor supply equation (e.g., a higher wage may induce women with higher tastes for leisure to enter the market) by including an inverse Mills' ratio term that is a function of the group/time participation rate. The labor supply equation that Blundell et al. (1998) actually estimate has the form

$$h_{it} = \beta \log w_{it} + \gamma[C_{it} - w_{it}(1-\tau_{it})h_{it}] + X_{it}\phi + d_g + d_t + \delta_w R_{wit} + \delta_c R_{cit} + M(P_{gt}) + \epsilon_{it}, \tag{115}$$

[80] An alternative computational approach to taking out group and time means is to regress the group mean of hours on the group mean of wages and a complete set of time and group dummies. Then the wage effect is identified purely from the wage variation not explained by time or group. The advantage of the more involved two-step procedure is that the coefficient on the residual provides a test of exogeneity of wages.

where τ_{it} is the tax rate, the second term is "virtual" non-labor income, X_{it} is a vector of demographic variables (for example, dummy variables for children of various ages), d_g and d_t are the group and time dummy variables, R_{wit} and R_{cit} are residuals from the first stage regressions of wages and virtual income on the group and time dummies, and $M(P_{gt})$ is the Mills' ratio used to correct for nonparticipation. The authors estimate this hours function by OLS.

To implement this procedure Blundell et al. (1998) group the FES data into 2 education groups (legal minimum vs. additional education) and 4 cohorts (people born in 1930-1939, 1940-1949, 1950-1959 and 1960-1969), or 8 groups in total. They include only 20 to 50 year old women with employed husbands. This gives 24,626 women of whom 16,781 work. Note that only workers are used to estimate (115), although the full sample is used to estimate the Mills' ratio. One detail is that 2970 of these women are within a few hours of a kink point in the tax schedule. Blundell et al. choose to drop these women from the data and construct an additional Mills' ratio term to deal with the selection bias this creates. They find that the group/time interactions are highly significant in the wage and virtual income equations.

The estimates imply an uncompensated wage elasticity at the mean of the data of 0.17 and a compensated elasticity of 0.20. In a sensitivity test, the authors report results where, in the first stage, the over-identifying instruments are 5 parameters that describe the tax rules interacted with group dummies. This reduces the number of instruments relative to the case where the group dummies were fully interacted with time dummies. It also means that only variation in wages and virtual income specifically induced by tax changes is used to identify the labor supply elasticities. The estimates give an uncompensated elasticity of 0.18 and an essentially zero income effect. Thus, results are little affected.

4.1.4. DCDP models

The first paper to adopt a full solution approach to modeling female labor supply was Eckstein and Wolpin (1989b). The main focus of the paper is on how the decision to work today affects wages and tastes for work in the future. Thus, the paper focuses on three of the four issues central to the female labor supply literature (i) fixed costs of working, (ii) human capital accumulation, and (iii) state dependence in tastes for work. To make estimation feasible (particularly given the 1989 computing technology) Eckstein and Wolpin (1989b) make some key simplifying assumptions. First, they ignore savings and assume a static budget constraint. Second, they ignore the choice of hours of work and treat labor supply as a discrete work/no-work decision.

This set of decisions is notable, as it illustrates well the different paths that the male and female life-cycle labor supply literatures have taken. The life-cycle literature on males has emphasized decisions about hours and savings, which Eckstein and Wolpin (1989b) ignore, while in most cases ignoring participation, human capital and state dependence, which they stress. This is not a value judgement on either literature, but simply an observation about what aspects of behavior researchers have found most essential to

model in each case. The emphasis on participation, human capital and state dependence explains why the female labor supply literature came to the use of DCDP models several years earlier than the male labor supply literature, as these features are very difficult to handle using Euler equation methods.

A third simplifying assumption that Eckstein and Wolpin (1989b) make is that they do not model marriage or fertility. To avoid having to model fertility decisions, the paper looks only at women who were at least 39 years old in 1967 (and hence for the most part past child bearing age). The number of children affects the fixed costs of work, but it is treated as a predetermined variable. Marriage is taken as exogenously given. Including marriage and fertility as additional choice variables would not have been feasible given 1989 technology, but, as we will see, incorporating them as choice variables has been the main thrust of the subsequent literature.

Eckstein and Wolpin (1989b) assume a utility function for married woman i at age t given by

$$U_{it} = C_{it} + \alpha_1 p_{it} + \alpha_2 C_{it} p_{it} + \alpha_3 X_{it} p_{it} + \alpha_4 N_{it} p_{it} + \alpha_5 S_i p_{it} \tag{116}$$

where p_{it} is an indicator for labor force participation, X_{it} is work experience (the sum of the lagged p_{it}'s), N_{it} is a vector of numbers of children in various age ranges (0-5 and 6-17) and S_i is the woman's completed schooling. The budget constraint is specified as

$$C_{it} = w_{it} p_{it} + y_t^H - cN_{it} - bp_t \tag{117}$$

where w_{it} is the wife's wage (annual earnings) if she works and y_t^H is the annual income of the husband (assumed exogenous).[81] The assumption that utility is linear in consumption has some important consequences. First, substitution of (117) into (116) makes clear that we cannot separately identify the fixed cost of work b and the monetary costs of children c from the disutility of work α_1 and the effect of children on the disutility of work α_4. Thus, b and c are normalized to zero.

The second implication of this specification is that the model will exhibit no income effects on labor supply unless consumption and participation interact in the utility function. If $\alpha_2 = 0$, then husband's income will have no impact on the wife's labor supply. A clear pattern in the data is that women with higher income husbands are less likely to work, which would imply that $\alpha_2 < 0$. Thus, to fit the data, consumption and leisure must be complements in utility, although in general, a negative income effect and consumption/leisure complementarity are conceptually distinct phenomena.

Eckstein and Wolpin (1989b) assume a standard log earnings function (linear in schooling, quadratic in work experience) with both a stochastic productivity shock and measurement error. A key point is that there are no shocks to tastes for work, so the only

[81] Annual earnings if the woman works are assumed to equal 2000 times the hourly wage rate, regardless of how many hours the woman actually works. This is necessitated by the 1/0 nature of the work decision.

stochastic components in the model are the productivity shocks and measurement error. This simplifies the solution to the dynamic programming problem.[82] The solution takes the form of a sequence of reservation wages (contingent on age, work experience and other state variables). The decision rule for participation is simply to work if the offer wage exceeds the reservation wage, which is a deterministic function of the state. The measurement error accounts for cases where women are observed to make decisions that violate this condition.

Eckstein and Wolpin (1989b) estimate the model by maximum likelihood using data on 318 white married women from the NLS Mature Women's cohort. The NLS interviewed them 11 times in the 16 years from 1967 to 1982, making it difficult to construct complete employment histories for all the women. To be in the sample, the women had to have at least four consecutive valid years of data on labor force participation and have a spouse present in every interview from 1967 to 1982. The data set contained 3020 total observations, 53% of which were for working years. The discount factor is fixed at 0.952.

An interesting aspect of the estimates is that they show substantial selection bias in OLS wage equation estimates. The OLS schooling coefficient is 0.08, while the model estimate (which corrects for selection) is 0.05. The experience profile is initially less steep but also less strongly concave than implied by OLS. The estimates also imply that 85% of observed wage variation is measurement error.[83]

With regard to the utility function estimates, Eckstein and Wolpin (1989a,b) find that children (especially young children) negatively affect tastes for work, as expected. The impact of state dependence is imprecisely estimated, but it implies that experience reduces tastes for work. Schooling reduces tastes for work as well. However, both taste effects are clearly outweighed by the positive effects of experience and schooling on wage offers.

Eckstein and Wolpin (1989b) find that $\alpha_2 < 0$; thus, as expected, husband income reduces the wife's participation rate. To quantify the size of the income effect, they consider a woman at age 39 with 15 years of work experience, 12 years of schooling, no children and a husband with $10,000 in annual earnings (which is close to the mean in the data). The baseline prediction of the model is that she will work 5.9 years out of the 21 years through age 59, or 28% of the time. If husband's earnings increase 50% the model predicts her participation rate will drop by half, to 14%. So the elasticity of the participation rate with respect to non-labor income is roughly 1.0. Converting this to an

[82] Eckstein and Wolpin (1989b) also assume that husband's earnings is a deterministic function of husband's age, a fixed effect, and a schooling/age interaction. If there were taste shocks or shocks to husband's earnings they would have to be integrated out in solving the DP problem.

[83] Note that the measurement error in wages cannot be estimated using wage data alone. But joint estimation of a wage equation and a labor supply model does allow measurement error to be estimated, as true wage variation affects behavior while measurement error does not. Of course, any estimate of the extent of measurement error so obtained will be contingent on the behavioral model.

income effect, and noting that the mean wage in the data is $2.27 dollars per hour and work is assumed to be 2000 hours per year, we obtain an income effect of −0.45.

Unfortunately, Eckstein and Wolpin (1989b) do not report a simulation of how an exogenous change in the wage rate (an increase in the intercept, the skill rental price, of the log wage function) would affect labor supply. However, as schooling is exogenous, and the effect of schooling on tastes for work is quantitatively small, we can approximate this using the estimated schooling coefficient. Consider, the same representative woman described above, and assume her education level is increased from 12 to 16. An extra 4 years of schooling raises the wage rate roughly 22% at the mean of the data. The model predicts that this will cause her participation rate from age 39 to 59 to increase by 108%. Thus, the implied (uncompensated) elasticity of the participation rate with respect to the wage is roughly 5.0.

Finally, Eckstein and Wolpin (1989b) report a detailed description of how the model fits labor force participation rates, conditional on 28 experience and age cells. In general, the model provides a very good fit to the data. As we noted earlier, there are very few papers in the static labor literature, or the literature on dynamic models based on first order conditions, that examine model fit. In contrast, the careful examination of model fit in the DCDP literature has become standard practice. The focus of the former literature is on estimation of parameters or elasticities, while the focus of the DCDP literature is on model simulations under baseline vs. counterfactual scenarios. It is only natural to compare the simulated baseline data to the actual data. Keane and Wolpin (2009) argue that it ought to be the industry standard to assess model fit in all econometric models (including static models, nonstructural models, etc.).

The next paper in the DCDP literature on female labor supply did not appear until Van der Klaauw (1996), which extended Eckstein and Wolpin (1989b) to include marriage as a choice. Women have up to 4 options in each period, given by the cross product of work and marriage choices. Another extension is that Van der Klaauw (1996) models decisions starting from when a woman has left school (rather than age 39, as in Eckstein and Wolpin), which may be as young as 14. Obviously then, he cannot treat fertility as given. Thus, Van der Klaauw (1996) models the arrival of children as a stochastic process, where arrival probabilities depend upon the state variables (i.e., marital status, education, age and race). This is a common practice in DCDP modeling—that is, to take variables that one believes are endogenous, but which one does not wish to model explicitly as a choice (either for computational reasons or because they are not the main focus of the analysis), and treat them as being generated by a stochastic process that depends on the other state variables.[84]

[84] This method is not necessarily more parsimonious than modeling a variable as a choice, trading off an additional choice variable (whether to have a child in this case with the corresponding utility and cost parameters) against additional parameters governing the stochastic outcome (the probability of having a child). A limitation of this method is that it does not allow for effects of contemporaneous shocks, for example a high wage draw for the female, on the probability of having a child.

The model is in many ways similar to Eckstein and Wolpin (1989b), again incorporating a static budget constraint and a utility function that is linear in consumption. Van der Klaauw specifies the utility function, conditional on the participation (p_t) and marriage choice (m_t), as

$$U_{pm,t} = \alpha_1 m_t + (\alpha_2 + \alpha_3 m_t) p_t + (\beta_1 + \beta_2 p_t + \beta_3 m_t) C_{pm,t} + \epsilon_{pm,t}. \quad (118)$$

Consumption is interacted with participation, as in Eckstein and Wolpin (1989b), which enables the model to explain why women work less if they have high income husbands. Tastes for marriage (α_1) are allowed to depend on demographics, children and lagged marriage. Marriage, m_{it}, is also interacted with consumption, C_{it}, thus letting marriage shift the marginal utility of consumption. The effects of demographics, children and lagged participation on tastes for work are captured by letting α_2 and α_3 depend on these variables. There is a separate taste shock for each of the mutually exclusive choices, $\epsilon_{mp,t} = (\epsilon_{11,t}, \epsilon_{10,t}, \epsilon_{01,t}, \epsilon_{00,t})$.

Recall that in Eckstein and Wolpin's (1989b) model a woman received utility from total household consumption. Here, a woman is assumed to consume her own income plus a fraction of the husband's income (which depends on her work status), so she receives utility from private consumption. A single women has a probability each year of receiving a marriage offer. The potential husband is characterized by his mean wage, which depends on the woman's characteristics (reflecting marriage market equilibrium) and a transitory wage draw.

It is worth noting that this is a search model of marriage only in a trivial sense. There is no match-specific component to the marriage. That is, a husband does not come with a permanent component to his earnings level, which could make him a "good draw" given the woman's demographics. Nor is there any permanent component to the utility level he provides. Thus, the woman has no reason to decline a marriage offer in the hope of a better offer. Her only reason for systematic delay is that mean husband income is found to be increasing in the woman's potential experience, and thus, her age. This setup substantially reduces the computational burden of estimation, as there is no "husband type" variable that must be included in the state space. But at the same time, the model is not informative about the effect of permanent differences in husband income on the wife's labor supply, as all permanent differences are a deterministic function of the wife's own characteristics.

The woman's own wage offer function includes standard covariates, such as education, a quadratic in experience, race, age and region. It also includes a lagged participation indicator, which allows recent work experience to be relatively more important. An unusual aspect of the specification, however, is that it is specified in levels, with an additive error. This is also true of the husband's wage function. The reason for adopting this specification is that, when these functions are substituted into the budget

constraint to obtain the choice-specific consumption level and this in turn is substituted into the utility function, each of the 4 alternatives turns out to have an additive error that consists of the relevant $\epsilon_{mp,t}$, plus a function of the female and male wage equation errors.

From a computational point of view, what enables handling the additional complexity of making marriage a choice is the assumption that these four additive choice-specific error terms, say $e_{mp,t}$ for $p = 0, 1, m = 0, 1$, are assumed to be distributed iid extreme value. As we have previously discussed, this assumption leads to closed form solutions for the DP problem and for the likelihood function. As also noted, the cost of making the extreme value assumption is that (i) it is contrary to the evidence suggesting that wage errors are approximately log normal and (ii) it assumes that shocks are contemporaneously uncorrelated. This latter assumption is very strong given that the four composite errors contain common error components; for example, $e_{01,t}$ and $e_{11,t}$ have husband income shocks in common.[85]

The model is estimated on PSID data from 1968 to 1985. The sample includes 548 females aged 12 to 19 in 1968 (29 to 36 in 1985), so that complete work and marital histories can be constructed (avoiding the initial conditions problem that would arise for women who were older in 1968). The terminal period is set at age 45 to reduce computational burden. It is assumed that $p_t = 1$ if the woman worked at least 775 hours in a year, but, as in Eckstein and Wolpin (1989a,b), the work choice is assumed to entail 2000 hours of work regardless of actual hours. An approximation is necessary due to the binary nature of the work decision.

The model is estimated in stages. In the first stage, the "reduced form" model with the woman's and the husband's wage equations substituted into (119) is estimated. In the second stage, the wage equations are estimated using employment and marriage decision rules from the reduced form model to implement a selection correction. In the third stage, a minimum distance estimator (see Chamberlain (1984)) is used to recover the structural parameters.

The estimates of the wage equations are a bit difficult to compare to prior literature as they are in levels. For instance, they imply that a year of schooling raises a woman's earnings by \$1379 per year. As mean earnings in the data are \$13,698 per year, this is roughly 10% at the mean of the data. A year of schooling also raises potential husband's earnings by \$1266 per year (vs. a mean of \$19,800) or 6.4%. This suggests that an important part of the return to schooling for women comes through the marriage market.[86] The utility function estimates imply that children reduce the utility from participation while lagged work increases the utility from participation.

[85] An alternative approach would be to assume the four errors follow a generalized extreme value distribution (see Arcidiacono (2005)).

[86] The estimates imply that a married woman who works receives 34% of husband income. Unfortunately, the share if she does not work is not identified. As can be seen from (118), if a married woman does not work her utility from consumption is $\beta_1 + \beta_3$ times her share of husband income. Only this product is identified in the model.

Van der Klaauw (1996) presents a substantial amount of evidence on the fit of the model, showing that it provides a good fit to the proportion of women who are working and married conditional on years since leaving school, to marriage rates by age, and to the hazard functions for marriage and divorce. It also provides a good fit to the proportion of women making each of the 4 marital status/work choices conditional on work experience and age.

Van der Klaauw (1996) then uses the model to simulate the impact of exogenous $1000 increases in annual offer wages and husband offer wages. The $1000 wage increase leads to a 26% (i.e., 2.5 year) increase in work experience by age 35. As this is a 7.3% wage increase, this implies an uncompensated labor supply elasticity of roughly 3.6. It is notable, however, that this elasticity is not comparable to a conventional Marshallian elasticity that holds all else fixed. In particular, the wage increase causes a 1 year increase in average years to first marriage, and a 1.3 year decrease in average total years of marriage. The reduction in marriage is part of what induces the increase in labor supply.[87]

The next significant paper in the DCDP literature on female labor supply is Francesconi (2002), which extends Eckstein and Wolpin (1989a,b) by making fertility a choice and allowing for both full- and part-time work. Thus, women have 6 choices in each annual period (after age 40 only the 3 work options are available). Francesconi (2002) also allows full and part-time experience to have separate effects on wage offers.[88] Thus, the model has three endogenous state variables: number of children, and part-time and full-time experience.

Marriage is taken to be exogenous and the model begins when a woman first gets married and ends at age 65. Women are assumed to make decisions based on the expected value of husband's income. As in Van der Klaauw (1996), the husband's mean income is purely a function of the woman's characteristics (i.e., age at marriage, education, education/age of marriage interactions, age). As in Eckstein and Wolpin (1989a,b) women receive utility from total consumption of the household, net of fixed costs of work and costs of children. There is again a static budget constraint, with utility linear in consumption. Utility for woman i at age t, conditional on her part-time and full-time work and fertility choices ($p_t \in (0, 1)$, $f_t \in (0, 1)$, $n_t \in (0, 1)$), is given by

$$U_{it} = C_{it} + \alpha_1 p_{it} + \alpha_2 f_{it} + (\alpha_3 + \epsilon_{it}^n) N_{it} + \alpha_4 N_{it}^2 + (\beta_1 p_{it} + \beta_2 f_{it} + \beta_3 n_{it}) C_{it} + (\beta_4 p_{it} + \beta_5 f_{it}) n_{it}. \quad (119)$$

87 Van der Klaauw (1996) simulates that a $1000 (or 5%) increase in husband offer wages would reduce average duration to first marriage by 1 year, increase average years of marriage (by age 35) by 2.3 years, and reduce average years of work by 2.6 years, or 27%. These are very large income effects, but they are not comparable to standard income effect measures, as they refer to changes in husband offer wages as opposed to changes in actual husband wages (or changes in some other type of non-labor income). Furthermore, it is not clear how much credence we can give to these figures since, as noted earlier, all permanent differences in husband income in the model are generated by differences in the wife's own characteristics.

88 There are separate part- and full-time wage functions.

The tastes for part and full-time work, α_1 and α_2, are allowed to be a function of the stock of children, N_{it}, work experience and schooling. Tastes for children vary stochastically over time, as captured by ϵ_{it}^n. Consumption is interacted with all the choice variables in order to allow husband's income to affect work and fertility decisions. Work and fertility decisions are interacted, which enables the model to capture the fact that women have lower participation rates during years that they have newborn children.

The stochastic terms in the model are the errors in the full and part-time log wage equations and the shock to tastes for children. There are no additional taste shocks. The errors are assumed to be distributed as joint normal. Thus, as in Eckstein and Wolpin (1989b), it is necessary to assume wages are measured with error to account for observations where women are observed to work at wages that are less than the reservation wage. Given that the model contains 6 choices and three error terms the evaluation of the Emax function integrals is difficult. Thus, Francesconi (2002) uses a simulation method like that proposed in Keane and Wolpin (1994) to evaluate the Emax functions. However, the state space is small enough that he can simulate the Emax function at every point in the state space (there is no need to interpolate between points). The three dimensional choice probability integrals are also simulated.

A point worth stressing is that Francesconi (2002) assumes that only the number of children, and not their ages, enters the state space. If children of different ages had different effects on labor supply, as we have previously noted, the size of the state space would grow astronomically. Francesconi can accommodate that newborns have a different effect on labor supply than older children, because newborns are treated as a current choice variable, and they do not enter the state (as they are no longer newborns in the next period). But allowing, e.g., the number of children aged 1 to 5 to have a different effect than the number of children aged 6-17, would greatly increase in complexity.

Francesconi (2002) also follows Van der Klaauw (1996) in limiting the size of the state space by assuming husband's mean income is purely a function of the woman's characteristics. Thus, husband-specific characteristics (e.g., a husband skill endowment) need not be included in the state space. Further, it is assumed that husband's earnings are realized only after the wife's labor supply and fertility decisions are made. As a result, the effect of husband's income on the wife's behavior can only be identified to the extent that there are exclusion restrictions, such that certain characteristics of the wife enter the model only through their effect on the husband's wage. In fact, the husband's wage function includes the wife's age, age at marriage and education/age of marriage interactions, and all of these variables are excluded from the wife's wage function and from her taste parameters.

Finally, Francesconi (2002) also extends earlier DCDP models of female labor supply by following the procedure in Keane and Wolpin (1997) to allow for unobserved heterogeneity. Specifically, he allows for three discrete types of women in terms of their

skill endowments (the intercepts in the offer wage functions) and in tastes for children (α_3 and α_4).

The model is estimated on a sample of 765 white women from the NLS Young Women Survey who were interviewed 16 times over the 24 years from 1968 to 1991. To be included in the sample the woman must be at least 19 and be continuously married to the same spouse during the sample period.[89] Part-time is defined as 500 to 1500 hours and full-time is defined as 1500+ hours. The discount factor is fixed at 0.952. In contrast to the multi-step procedure in Van der Klaauw (1996), the decision rules and wage offer functions are estimated jointly. There are separate wage offer functions for part-time and full-time work.

The estimates of the wage function imply that a year of schooling raises the full-time offer wage by 8.4% and the part-time offer wage by 7.6%, estimates that are intermediate between the Eckstein and Wolpin (1989b) and Van der Klaauw (1996) results. Full-time experience has a larger positive effect on full-time offer wages than part-time experience. Effects of experience on part-time offer wages are generally much smaller. Measurement error accounts for about 63% of the variance of observed wages. Evaluated at the mean of the data, an extra year of school raises mean husband wages by 11%. This is consistent with the finding of Van der Klaauw (1996) that a large part of the return to schooling for women comes through the marriage market rather than the labor market. The interaction terms between consumption and work and fertility (β_1, β_2, β_3) are all negative, which generates negative income effects on both labor supply and fertility. In addition, individuals of the type with a high skill endowment have relatively low tastes for children.

Francesconi (2002) reports results indicating that the model provides a good fit to all 6 annual choice options up to 24 years after marriage, which corresponds to age 47 on average (the last observed age in the NLSY79 data he analyzed). He also fits a static model (i.e., a model with the discount factor set to 0) and finds that it too provides a good fit to the in-sample data. But the models differ dramatically in their out-of-sample predictions. The static model predicts that women's labor supply will increase sharply after about age 47 and into their 60's. The DCDP model implies that work will stay flat and then drop slowly in their 60's. The latter prediction is much closer to what is observed in CPS data, which covers adult women of all ages.[90] The static model explains low participation rates as resulting from the presence of children; when children leave the household, participation rates rise sharply. In the dynamic model, the return to human capital investment, that is of working, falls as one approaches the terminal period, which counteracts the effect of children leaving.

[89] This is a sub-sample of a group of 1,783 women who were married at least once during the period (the larger sample including women who leave a partner during the sample period).

[90] Neither model captures the sharp decline in participation in their 60's due to retirement. But to be fair neither model incorporates any features designed to explain retirement behavior (such as pensions or Social Security).

Finally, Francesconi (2002) conducts a number of simulations of how permanent changes in wages would affect labor supply. For example, consider an average woman with 2 years of full-time work experience at the time of marriage. The baseline model simulation shows that she will work for 6.8 out of the 11 years from age 30 to 40. An increase in the log wage function intercept (which represents the rental price of skill) would increase offer wages at the mean of the data by roughly 10.5%, and it would increase full-time work by roughly 60%. This implies an elasticity of labor supply with respect to rental price of skill of roughly 5.6. However, this is somewhat of an exaggeration, as some of the increase in full-time work must come from reduced part-time work. Unfortunately, Francesconi (2002) does not report the decrease in part-time work that accompanies the increase in full-time work.

The last two papers on female labor supply described below are Keane and Wolpin (2007, 2010). In these papers, Keane and Wolpin utilize approximate solution methods developed in Keane and Wolpin (1994), and estimation methods developed in Keane and Wolpin (2001), to estimate a model of female life-cycle behavior that is considerably richer than previous models in the literature. Both marriage and fertility are treated as choices, and both full and part-time work options are available. Schooling is also a choice. An important feature of the data that is not accommodated in prior dynamic models is that a large fraction of single women with children participate in public welfare programs. Thus, welfare participation (when eligible) is also incorporated as a choice.

In the model, women begin making decisions at age 14, and the terminal period is age 65. The fertile period is assumed to last up until age 45, and during this period women have up to 36 choice options in each period. Afterwards they have up to 18 options.[91] The decision period is assumed to be 6-months until age 45, which is a compromise between the length of a school semester and the child gestation period. After age 45, the decision period is one year (as the fraction of women who either attend school or have children after 45 is negligible). Given that behavior of girls as young as 14 is being modeled, it is essential to consider the role of parental co-residence and parental income support. Yet, as this is not a focal point of the model, the authors choose not to treat living with parents as a choice. Both the probability of co-residence and parental transfers are treated as stochastic processes that depend on a person's state variables.

One fundamental difference from Van der Klaauw (1996) and Francesconi (2002) is that marriage is treated as a true search process. Each period a woman may receive a marriage offer that consists of: (1) the mean wage of the husband, and (2) a marriage quality draw (which captures nonpecuniary aspects of the match). The potential husband's mean wage depends on the woman's characteristics, such as her schooling and skill level, as well as a permanent component drawn from a distribution. Thus, a husband

[91] The choice set differs across women for a number of reasons. For instance, only unmarried women with children under 18 have the option to participate in welfare, and working while on welfare is not an option if the offer wage rate is high enough that income would exceed the eligibility level. Also, girls under 16 cannot choose marriage.

fixed effect becomes part of the state space. In this setup, a woman has an incentive to reject marriage offers while waiting for a husband with a high mean wage.

Another fundamental difference from prior work is that the model is non-stationary in the sense that the economic environment changes over time. Specifically, the welfare rules change over time and differ by state, so each cohort of women (as defined by the semi-annual period in which they reach age 14) in each state faces a different sequence of welfare rules. This creates a number of computational problems. First, each cohort of women in each state faces a different dynamic optimization problem (raising computational burden). Second, one must make an assumption about how women forecast future rules. Third, the rules are complex, making it difficult to characterize them.

Keane and Wolpin (2007, 2010) deal with these problems as follows. First, they develop a simple 5 parameter function that characterizes the welfare benefit rules in each state in each year quite accurately. Second, they assume women use a state-specific VAR in these 5 parameters to predict future rules. Third, they only use data from 5 large states, so as to reduce the number of DP problems that must be solved in estimation. This enables them to use the data from other states for out-of-sample validation.

Keane and Wolpin (2007, 2010) assume that a woman receives disutility from a variable that measures "non-leisure" time. This is a sum of work hours, a fixed time cost of work, time spent in school, time required to collect welfare, and time required to care for children.[92] The authors estimate weights on the variables other than work hours to account for the fact that school time, child care time and time collecting welfare may entail more/less disutility than time spent working. A woman receives utility from consumption, which is assumed to be a share of total household income. Utility is quadratic in non-leisure time and linear in consumption. Similar to the previous papers we discussed, consumption is interacted with non-leisure time. The estimated coefficient is negative, implying that consumption and leisure are complements, inducing negative income effects on labor supply and fertility.

Additional interactions are introduced that allow marriage and children to shift the degree of complementarity between consumption and leisure. This would have been irrelevant in the papers discussed previously, as they do not try to explain labor supply, marriage and fertility choices jointly. The estimates imply that marriage and children both significantly reduce the degree of complementarity between consumption and leisure, but do not eliminate it.

Women also receive utility/disutility from children, pregnancy, marriage, school attendance and welfare participation. Utility is quadratic in number of children. The utility/disutility from pregnancy is a polynomial in age. As one would expect, this becomes a large negative for women as they approach 45, consistent with the greater

[92] Childcare time is, in turn, a weighted sum of time required to care for children in different age ranges.

risks associated with pregnancy at older ages. The disutility of welfare attendance enables the model to explain the common phenomenon of nonparticipation by eligible women (see Moffitt (1983)). The utility function coefficient on each of the 5 choice variables (hours, pregnancy, marriage, school and welfare) consists of a constant plus a stochastic taste shock. This enables the model to generate a nonzero probability of any observed choice outcome.

The model allows for unobserved heterogeneity in the form of 6 types of women who differ in the preference parameters (constant terms) associated with the 5 choice variables (i.e., different tastes), and in the intercepts of the own and potential husband offer wage functions (i.e., different skills). The model includes observed heterogeneity as well; the heterogeneous skill and taste parameters differ across states and across ethnic groups (blacks, whites and Hispanics). Finally, the utility function includes interactions of indicators for full and part-time work, school and marriage with lagged values of these indicators, to capture state dependence in tastes for these choice options.[93]

The model is estimated using data from the National Longitudinal Survey of Youth 1979 cohort (NLSY79). The NLSY79 includes women aged 14 to 21 in 1979. The paper uses the data from the years 1979 to 1991. Thus, the women reach a maximum age of 33. The states used in estimation are California, Michigan, New York, North Carolina and Ohio. To be in the sample, a woman had to reside in the same state for the whole sample period, which screens out about 30%. This leaves data on approximately 2800 women.[94] The annual discount factor is fixed at 0.93.

Estimates of the log wage function imply that (at the mean of the data) an additional year of school raises wages by 9.1%. And 84% of the variance of wages is attributed to measurement error (the true log wage standard deviation is 0.17). The experience coefficients imply that the first year of full-time work raises wages by 2.6%, and that the experience profile peaks at 36 years. In addition, lagged full-time work raises the current wage offer by 7%, while lagged part-time raises it by 3%. Black and Hispanic women have lower offer wages than white women (by 13% and 6%, respectively).

In the husband offer wage function, the coefficient of the woman's skill endowment (i.e., intercept in the woman's wage function) is 1.95, implying a very high degree of assortative mating on skill. And each additional year of education for the woman raises the husband offer wage by 3%. Black and Hispanic women have much lower

[93] The utility function includes some miscellaneous additional terms that were added to capture some specific features of the data. Full and part-time work are interacted with school to capture the fact that people who work while in school tend to have a strong preference for part-time over full-time work. Work variables are also interacted with a school less than 12 dummy to capture that part-time work is far more prevalent among high school students. Pregnancy is interacted with school to capture that women rarely go to school while pregnant. Tastes for school, marriage and pregnancy are also allowed to shift at certain key ages (16, 18 and 21). And there is a linear time trend (across cohorts) in tastes for marriage.

[94] Keane and Wolpin (2002), which presents a nonstructural analysis of the same data, provides a more detailed description.

husband offer wages than whites (by 30% and 14%, respectively). The estimates imply that women receive 55% of total household income. So, just as in Van der Klaauw (1996) and Francesconi (2002), much of the return to schooling appears to emerge through the marriage market.

Keane and Wolpin (2007) provide a good deal of evidence on the fit of the model and assess how well it predicts behavior in the holdout state of Texas. The model performs reasonably well in these tests, including providing better predictions than some candidate competing nonstructural models.

As has been the focus of the labor supply literature, Keane and Wolpin (2010) estimate labor supply wage elasticities. Recall that the model has six types of women, which we can rank by skill level from type 1 (highest skill endowment) to types 6 (lowest). Type 6 account for the majority of welfare participants. Keane and Wolpin (2010) report experiments where they increase the offer wage by 5% for each type separately. The wage elasticities are inversely proportional to skill level, ranging from only 0.6 for type 1 to 9.2 for type 6. Thus, the overall elasticity of 2.8 is deceptive with regard to behavior of various subsets of the population.

For type 6 women, the 5% wage increase has a dramatic impact on all aspects of their behavior. For instance, for white women of type 6, the percent working at ages 22 to 29.5 increases from 34% to 50% (a 47% increase). But it is also notable that mean completed schooling increases from 11.5 to 12 years, the high school drop out rate drops from 42% to 24%, welfare participation drops from 25% to 20%, and incidence of out-of-wedlock teenage pregnancies drops from 3.4% to 2.8%. All of these behavioral changes (i.e., more education, fewer teenage pregnancies, less welfare participation) contribute to the increase in labor supply. In contrast, type 1 are already completing a high level of schooling, are rarely having children at young ages, are not participating in welfare, and are participating in the labor market at a high rate. Thus, in a sense there are fewer channels through which a wage increase can affect them. In summary, the results indicate that wage elasticities of labor supply for low skilled women are much greater than for high skilled women.

It is difficult to summarize the estimates of labor supply elasticities for women across the studies we have surveyed. Several of the non-DCDP studies we have examined calculate what might be called "short run" elasticities that hold work experience, marriage and fertility fixed. On the other hand, the DCDP models calculate "long run" elasticities that allow, depending on the study, some combination of experience, fertility, marriage and education to adjust to wage changes. Nevertheless, a reasonable assessment of the estimates from this literature is that the labor supply elasticity estimates for women are generally quite large. The DCDP models give uniformly large "long run" elasticities ranging from 2.8 to 5.6. The life-cycle models of Heckman and MaCurdy (1982) and Kimmel and Knieser (1998) give large Frisch elasticities (2.35 to 3.05). The Marshallian elasticity of 0.89 obtained by Cogan (1981) in a static model is

also quite large.[95] Thus, 7 of the 9 studies obtain large female labor supply elasticities (of various types). Only the Blundell and Walker (1986) and Blundell et al. (1998) studies find small elasticities. This may be because these two studies consider the labor response of working women to wage changes, while the other 7 studies incorporate the participation margin.

The richness of the Keane and Wolpin (2010) model enables them to address a variety of substantive issues beyond calculating labor supply elasticities. These focus on (i) the factors that account for differences between blacks, whites and Hispanics in choice behavior and (ii) the effects of changing welfare rules. With respect to behavioral differences among minority and white women, the model estimates indicate that black women face a worse marriage market than do white women. The mean earnings of potential husbands, that is, the pool of men who make marriage offers, is 27 percent lower for black than for white women. In addition, unobservable traits of potential mates reduce the psychic value of getting married by $2,500 (in 1987 NY dollars) for black women relative to white women. The estimates also indicate that black women face poorer labor market opportunities. Wage offers are 12.5 percent lower for black women than for white women. In terms of preferences, the stigma attached to being on welfare is smaller for black women, although the difference, 290 dollars per six month period, does not seem that large. Black women do not differ from white women in the disutility they attach to work (an extra 1000 hours of work is equivalent to a 117 dollar greater drop in consumption for black than for white women), but they are estimated to have a significantly greater preference for children (the birth of a child is equivalent to a greater increase in consumption by 1352 dollars for black women than for white women).

To assess the importance of labor market, marriage market and preference differences, Keane and Wolpin (2010) simulate behaviors of black women under alternative counterfactual scenarios. They find that equalizing marriage market opportunities between black and white women would reduce welfare participation of black women at, for example, ages 26-29 from 29.7 percent to 21.4 percent, thus closing 37 percent of the black-white gap. Equalizing labor market opportunities has a somewhat larger impact, reducing the gap by about 45 percent. However, these changes have opposite effects on employment. Improving marriage market opportunities of black women in this age group reduces their employment rate from 55.7 to 42.3 percent, and thus widens the black-white gap, while employment rates are essentially equalized when labor market opportunities are equalized. Both counterfactuals increase marriage rates in that age range, although directly operating on marriage market conditions has a much larger impact, reducing the marriage rate gap of 37 percentage points to only 10 percentage points. Along with this large increase in marriage rates, the mean number of teenage births increase slightly. On the other hand, the relatively small increase in marriage

[95] Note that this is an elasticity for hours conditional on working. It is unfortunate that Cogan does not report a participation elasticity, as, given his estimates, this would presumably have been much larger.

rates that accompany the counterfactual improvement in the labor market leads to a fall in the mean number of teenage births by 13 percent and closes the black-white gap by 38 percent. Finally, improving the marriage market of black women reduces their completed schooling by a third of a year on average, widening the gap with white women, while improving their labor market opportunities increases their completed schooling by 0.2 years. Increasing welfare stigma of black women to that of white women, given the relatively small difference noted above, has only a small impact on behavior; the largest effect is to reduce welfare participation by 3.3 percentage points, significantly less than that exhibited for the other counterfactuals.

As these counterfactual experiments illustrate, none of these differences, when taken one at a time, can account for the racial differences in outcomes. Improving marriage market opportunities, by itself, reduces some of the gaps, but widens others. Improving labor market opportunities reduces all of the gaps, but considerably less so for demographic outcomes. And, welfare stigma accounts for little of the racial differences in behavior.

In another counterfactual experiment, Keane and Wolpin (2010) simulate the effect of eliminating welfare. Because welfare receipt is heavily concentrated among one of the six (unobserved) types, this experiment was performed only for women with the preferences and opportunities of the women of that type. For this type, 68.1 percent of black women and 24.6 percent of white women are receiving welfare at ages 26-29. That difference, of 43.5 percentage points is eliminated in the experiment. But, perhaps the most striking result is that eliminating welfare also essentially eliminates the employment gap, even though labor market opportunities are worse for black women. The original gap of 16.5 percentage points is reduced to 1.4 percentage points. Eliminating welfare also increases marriage rates more for black women, by 14.8 percentage points, than for white women, 8.2 percentage points, reducing the original gap from 36.2 percentage points to 29.6 percentage points. The mean number of teenage births fall slightly, but about the same for black and white women. A similar result is observed for the proportion of women of this type who do not graduate from high school.

As Keane and Wolpin (2010) conclude, there is no simple answer as to what causes the differences in the behavior of black and white women. The welfare system in place in the US until the major reform in 1996 differentially affected the labor market attachment of black women, but did not by itself account for much of the difference in marriage, fertility and schooling. The poorer marriage and labor market opportunities of black women both contributed importantly to the greater dependency of black women on welfare. Ultimately, it is the interaction of all of these factors, the welfare system, opportunities and preferences that jointly account for the large racial gaps in labor market and demographic outcomes.

In summary, the female labor supply literature has emphasized the connection between participation decisions and human capital, fertility and marriage. Those papers

that have attempted to model fertility and/or marriage as choices have ignored savings behavior to achieve computational tractability. There is as of yet no model of female life cycle behavior that includes savings along with human capital, fertility and marriage. This is an important avenue for future research, although a difficult one, because it involves modeling interactions within a household in a dynamic framework.[96]

4.1.5. Male labor supply

As we have noted, DCDP models of female labor supply have ignored considerations of consumption smoothing through savings and borrowing behavior. This is in sharp contrast to the literature on male labor supply, which has made consumption smoothing a major focus, in conjunction with a continuous hours choice and, with few exceptions, has ignored human capital accumulation.

Indeed, without availing itself of the DCDP approach, the literature on males has generally adopted estimation methodologies that specifically seek to avoid having to solve the full dynamic programming problem. A notable example is the seminal work by MaCurdy (1981, 1983), who developed estimation methods using the Euler conditions of dynamic models of labor supply with savings. Shaw (1989) extended this approach to a model with labor supply, savings and human capital accumulation. Of course, the DCDP methodology does not preclude modeling all these aspects of behavior, but it is computationally burdensome. On the other hand, a limitation of Euler equation and other "non-DCDP" or "non-full solution" approaches, is that, while they can deliver structural parameter estimates, they do not in general allow one to simulate behavioral responses to changes in policy (or the economic environment more generally).

To our knowledge only one paper, Imai and Keane (2004), has used DCDP methodology to estimate a labor supply model with assets, continuous hours and on-the-job human capital accumulation. We present a simplified version of the Imai and Keane (2004) model that captures the main points.

Assume that a worker's human capital, denoted by K, evolves according to the simple human capital production function

$$K_{i,t+1} = (1 + \alpha h_{it})K_{it}. \tag{120}$$

The growth in human capital, in this formulation, is a constant fraction of hours worked. K_{i1} is the person's skill (or human capital) endowment at the time of labor force entry.[97] A person's wage at time t, w_{it}, is equal to the current stock of human capital times the (constant) rental price of human capital, R.[98] Human capital is subject to a

[96] For recent attempts see Mazzocco and Yamaguchi (2006) and Tartari (2007).

[97] Imai and Keane (2004) actually assume a much more complex process, designed to capture patterns of complementarity between human capital and hours of work in the human capital production function. But use of this simpler form helps to clarify the key points.

[98] Although Imai and Keane assume a constant rental price, allowing for time varying rental rates is fairly straightforward.

transitory productivity shock. Specifically,

$$w_{it} = RK_{it}(1 + \epsilon_{it}). \tag{121}$$

The period-specific utility function is given by

$$U_t = \frac{C_t^{1+\eta}}{1+\eta} + \beta_t \frac{h_t^{1+\gamma}}{1+\gamma}, \quad \eta \leq 0, \gamma \geq 0, \tag{122}$$

where β_t is an age-varying parameter that shifts tastes for work. In contrast to the female labor supply literature, where utility is typically assumed to be linear in consumption, utility in (122) is CRRA in consumption. Given the emphasis of the male literature on savings, the CRRA is a more natural choice. Assets evolve according to

$$A_{t+1} = (1 + r)(A_t + (1 - \tau)w_t h_t - C_t) \tag{123}$$

where τ is the tax rate on labor income. Given this setup, the state space at t, Ω_t, consists of $\{K_t, A_t, \epsilon_t, \beta_t\}$.

The individual is assumed to maximize the expected present discounted value of utility over a finite horizon. The value function at age t is then

$$V_t(\Omega_t) = \max_{C_t, h_t} U_t + E_t \left(\sum_{j=t+1}^{T} \delta^{j-t} U_j \right) \tag{124}$$

$$= \max_{C_t, h_t} U_t + \delta E(V_{t+1}(\Omega_{t+1}) | \Omega_t, C_t, h_t), \tag{125}$$

where the expectation is taken over the future transitory productivity shocks and tastes for work conditional on the current state space. As is common in these types of models, we assume these stochastic terms are independent over time. In that case, we can replace Ω_t with Ω_t^-, that is, we can drop ϵ_{it} and β_t from Ω_t in forming the expectation in (125). Note that $E(V_{t+1}(\Omega_{t+1}^-, \epsilon_{t+1}, \beta_{t+1}) | \Omega_t^-, C_t, h_t)$ is simply the analog of the $E \max_{t+1}$ function in the discrete choice problem already discussed.

As in the discrete choice case, the solution of the model consists of finding the $E \max_t$ functions. Imai and Keane (2004) do that using a backsolving and approximation procedure similar to Keane and Wolpin (1994), adapted to continuous choice variables. In the terminal period, the value function is

$$V_T(\Omega_T) = \max_{C_T, h_T} \left(\frac{C_T^{1+\eta}}{1+\eta} + \beta_T \frac{h_T^{1+\gamma}}{1+\gamma} \right). \tag{126}$$

In this simple static problem, and without a bequest motive, given w_T and A_T, the consumer chooses C_T and h_T to maximize utility subject to the budget constraint $C_T = (1-\tau)w_T h_T + A_T$.[99]

In principle, the backsolving procedure starts by calculating $V_T(\Omega_T)$ for every possible state in Ω_T at which the worker might enter period T. The solution for h_T is given by the first-order condition

$$\frac{\beta_T h_T^{\gamma}}{[(1-\tau)w_T h_T + A_T]^{\eta}} = (1-\tau)w_T. \tag{127}$$

This equation can be solved numerically for the optimal h_T using an iterative search procedure. Once the optimal h_T is determined for each state point, the optimal C_T is found from the budget constraint. $V_T(\Omega_T)$ is then found by substituting the optimal value of h_T and C_T into (126).

Although we need to calculate $E\max_T$ only at the deterministic components of the state space, Ω_T^-, a problem arises, because the number of possible levels of human capital and assets at the start of period T is extremely large, if not infinite. Thus, it is not computationally feasible to literally solve for $E\max_T$ for every possible state value. Thus, Imai and Keane (2004) adopt the Keane and Wolpin (1994) approximation method discussed earlier, which involves solving for $E\max_T$ at a finite (and relatively small) subset of the possible state points.

To implement that procedure, a regression is estimated as some flexible function of the state variables and used to predict or interpolate the value of $E\max_T$ at any desired state point (K_T, A_T), including, in particular, points that were not among those used to fit the regression. Thus, having fit this interpolating regression, we may proceed as if $E\max_T$ is known for every possible state point in (K_T, A_T). As before, denote the interpolating function that approximates $E\max_T$ as $\widehat{E\max_T}$. We must assume that $\widehat{E\max_T}$ is a smooth differentiable function of K_T and A_T (e.g., a polynomial) for the next step. For expositional convenience, let $\widehat{E\max_T}$ be the following simple function,

$$\widehat{E\max_T} = \pi_{T0} + \pi_{T1}\log K_T + \pi_{T2}\log A_T + \varsigma_T, \tag{128}$$

and let $\pi_T(K_T, A_T) = \widehat{\pi}_{T0} + \widehat{\pi}_{T1}\log K_T + \widehat{\pi}_{T2}\log A_T$ be the predicted value of $\widehat{E\max_T}$, where the $\widehat{\pi}$'s are estimated parameters.

As in the discrete choice setting, the next step of the backsolving process moves back to period $T-1$. Then, using the predicted values of $E\max_T$ from the approximating

[99] Adding a bequest motive to the model, as in Imai and Keane (2004) is straightforward. This extension can be accommodated by adding a terminal value function, say $f(A_{T+1})$ to $V_T(\Omega_T)$.

function,

$$V_{T-1}(\Omega_{t-1}) = \max_{C_{T-1},h_{T-1}} U_{T-1} + \delta\pi_T(K_T, A_T). \tag{129}$$

Upon substituting in the laws of motion for K_T and A_T, we get

$$\begin{aligned} V_{T-1}(\Omega_{t-1}) = \max_{C_{T-1},h_{T-1}} & \frac{C_{T-1}^{1+\eta}}{1+\eta} + \beta_{T-1}\frac{h_{T-1}^{1+\gamma}}{1+\gamma} \\ & + \delta(\widehat{\pi}_{T0} + \widehat{\pi}_{T1}\log((1+\alpha h_{i,T-1})K_{i,T-1}) \\ & + \widehat{\pi}_{T2}\log((1+r)(A_{T-1} + (1-\tau)w_{T-1}h_{T-1} - C_{T-1}))). \end{aligned} \tag{130}$$

Finding the optimal values of consumption and hours is now just like a static optimization problem. The first order conditions are given by

$$\begin{aligned} \frac{\partial V_{T-1}}{\partial C_{T-1}} &= C_{T-1}^{\eta} - \frac{\delta\widehat{\pi}_{T2}}{A_{T-1} + (1-\tau)w_{T-1}h_{T-1} - C_{T-1}} = 0, \\ \frac{\partial V_{T-1}}{\partial h_{T-1}} &= -\beta_{T-1}h_{T-1}^{\gamma} + \frac{\delta\alpha\widehat{\pi}_{T1}}{1+\alpha h_{T-1}} + \frac{\delta(1-\tau)w_{T-1}\widehat{\pi}_{T2}}{A_{T-1} + (1-\tau)w_{T-1}h_{T-1} - C_{T-1}} = 0. \end{aligned} \tag{131}$$

These two equations can be solved numerically for C_{T-1} and h_{T-1} at any given state point in Ω_{t-1}.[100]

Following the development for period T, the next step is to calculate the values of $E\max_{T-1}$ at a subset of the state points. For a given value of K_{T-1} and A_{T-1}, we can substitute the optimal C_{T-1} and h_{T-1}into V_{T-1} (130) and numerically integrate over the joint distribution of ϵ and β. Given the values of $E\max_{T-1}$, we can then estimate the interpolating function at $T-1$, say

$$\widehat{E\max}_{T-1} = \pi_{T-1,0} + \pi_{T-1,1}\log K_{T-1} + \pi_{T-1,2}\log A_{T-1} + \varsigma_{T-1}. \tag{132}$$

Using this interpolating function, we can write the (approximate) value functions at time $T-2$ in an analogous fashion to (130). The only difference is in the interpolating function parameters. These steps are repeated until an approximate solution is obtained for every period back to $t = 1$.

The approximate solution consists of the complete set of interpolating function parameters, the $\widehat{\pi}_t$'s for $t = 2, \ldots, T$. Given these estimated interpolating functions, it is possible to solve numerically the simple two equation system like (131) at each t to

[100] These equations may have multiple solutions. If there are, then one would need to check second order conditions or calculate V_{T-1}.

find the optimal choice of a worker at any point in the state space. In particular, using $\widehat{E\max}_2$, one can solve for optimal labor supply and consumption in period $t = 1$, the first period of the working life. As previously noted, this is what first order conditions alone do not provide. Furthermore, by drawing values for the taste shocks and rental rates and repeatedly solving optimal labor supply and consumption over time, one can simulate entire career paths of workers. This enables one to simulate how changes in the economic environment, such as changes in tax rates, would affect the entire life-cycle path of labor supply and consumption, as one can re-solve the model and simulate career paths under different settings for the policy parameters.

Imai and Keane (2004) estimate their model using white males from the NLSY79. They choose this data set because of its fairly extensive asset data. The men in their sample are aged 20 to 36 and, as the focus of their paper is solely on labor supply, they are required to have finished school. Due to the computational burden of estimation they randomly choose 1000 men from the NLSY79 sample to use in estimation. People are observed for an average of 7.5 years, each starting from the age at school completion.

Notably, Imai and Keane (2004) allow for measurement error in observed hours, earnings and assets when constructing the likelihood function. As all outcomes are measured with error, construction of the likelihood is fairly simple. One can simulate career histories for each worker, and then form the likelihood of a worker's observed history of hours, earnings and assets as the joint density of the set of measurement errors necessary to reconcile the observed history with the simulated data.[101]

Imai and Keane (2004) estimate that $\gamma = 0.26$. In a model without human capital, this would yield a Frisch elasticity of $\frac{1}{\gamma} = 3.8$, which implies a much higher willingness to substitute labor intertemporally than in almost all prior studies for men (see MaCurdy (1983) for an exception). Simulations of the model reveal that, even accounting for human capital effects, the estimate of γ implies more elastic labor supply than in most prior work.

Imai and Keane (2004) explain their high estimate of intertemporal substitution based on the logic of Fig. 1. The figure presents a stylized (but fairly accurate) picture of how wages and hours move over the life cycle. Both wages and hours have a hump shape, but the hump in wages is much more pronounced. This apparently weak response of hours to wages leads conventional methods of estimating the intertemporal elasticity of substitution (which ignore the effect of working on the accumulation of human capital) to produce small values.

Indeed, Imai and Keane (2004) show that if they simulate data from their model and apply instrumental variable methods like those in MaCurdy (1981) and Altonji (1986) to estimate $\frac{1}{\gamma}$, they obtain values of 0.325 (standard error = 0.256) and 0.476 (standard

[101] Keane and Wolpin (2001) first developed this approach to forming the likelihood in DCDP models. Keane and Sauer (2009) extended the approach to nonstructural panel data models.

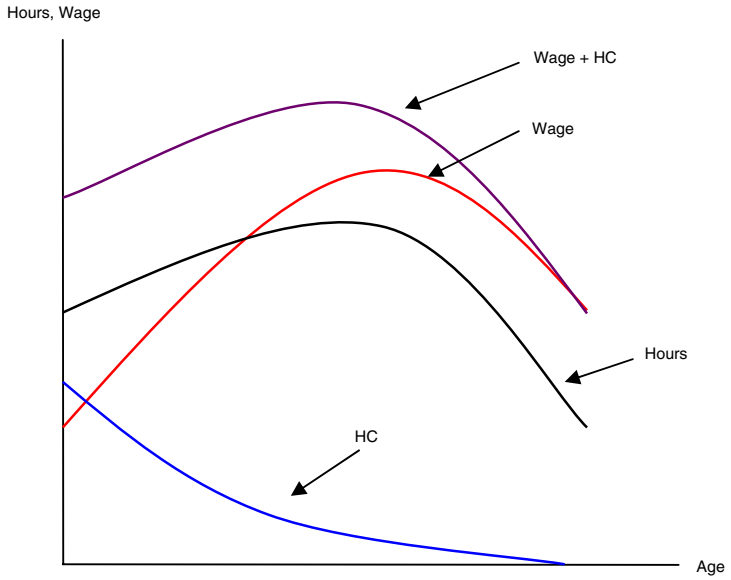


Figure 1 ***Hours, wages and price of time over the life-cycle.*** Note: HC denotes the return to an hour of work experience, in terms of increased present value of future wages. The opportunity cost of time is Wage + HC.

error = 0.182), respectively. Thus, the Imai and Keane (2004) model generates life-cycle histories that, when viewed through the lens of models that ignore human capital accumulation, imply similarly low co-movement between hours and wages to those obtained in most prior work. As further confirmation of this point, the authors report simple OLS regressions of hours changes on wage changes for both the NLSY79 data and the data simulated from their model. The estimates are −0.231 and −0.293, respectively. Thus, a negative correlation between hours changes and wage changes in the raw data is perfectly consistent with a high willingness to substitute labor intertemporally over the life cycle.

What reconciles these prima facie contradictory observations is the divergence between the opportunity cost of time and the wage in a model with returns to work experience. In particular, Imai and Keane (2004) estimate that from age 20 to 36 the mean of the opportunity cost of time increases by only 13%. In contrast, the mean wage rate increases by 90% in the actual data, and 86% in the simulated data. Thus, the wage increases about 6.5 times faster than the opportunity cost of time. These figures imply that conventional methods of calculating $\frac{1}{\gamma}$ will understate it by a factor of roughly 6.5.

This point is illustrated in Fig. 1 by the line labeled "Wage + HC," which adds the wage and the return to an hour of work experience (in the form of higher future earnings) to obtain the opportunity cost of time. As the figure illustrates, the opportunity

Table 2 Effect of a 5% tax on earnings on labor supply by age.

Age	Pure tax	Tax plus lump sum redistribution
20	−0.7%	−3.2%
30	−0.7%	−3.3%
40	−0.9%	−4.2%
45	−1.2%	−5.7%
50	−2.1%	−8.7%
60	−9.1%	−20.0%
20-65 (Total hours)	−2.0%	−6.6%

cost of time is much flatter over the life cycle than is the wage rate. Thus, hours appear to be much more responsive to changes in the opportunity cost of time than to changes in wages alone.

Imai and Keane (2004) use their model to simulate how workers of different ages would respond to a 2% temporary unanticipated annual wage increase. For a worker at age 20, hours increase only 0.6%. But the response grows steadily with age. At age 60 the increase in hours is nearly 4%, and at age 65 it is about 5.5%. The reason the effect of a temporary wage increase rises with age is that, as depicted in Fig. 1, as a person ages the current wage becomes a larger fraction of the opportunity cost of time. According to Imai and Keane (2004)'s estimates, at age 20 the wage is less than half of the opportunity cost of time, but by age 40 the wage is 84% of the opportunity cost of time.

Unfortunately, the Imai and Keane (2004) simulations do not reveal what the model implies about how workers would respond to permanent tax changes. To fill this gap, Keane (2009a) uses the Imai-Keane model to simulate the impact of a permanent 10% tax rate increase (starting at age 20 and lasting through age 65) on labor supply over the entire working life. If the tax revenue is simply thrown away, the model implies that average hours of work (from ages 20 to 65) drops from 1992 per year to 1954 per year, a 2% drop. If the revenue is redistributed as a lump sum transfer, labor supply drops to 1861 hours per year, a 6.6% drop. The later is as a reasonable approximation to the compensated elasticity with respect to permanent tax changes implied by the model (i.e., 1.32).

The effects of the tax, however, are very different at different ages. As seen in Table 2, tax effects on labor supply slowly rise from age 20 to about age 40. Starting in their 40's, the effects on labor supply start to grow quite quickly, and by age 60 effects are substantial. Thus, in response to a permanent tax increase, workers not only reduce labor supply, but also shift their lifetime labor supply out of older ages towards younger ages.

To our knowledge, there are only two papers besides Imai and Keane (2004) that have used full solution methods to estimate a life-cycle model that includes both human capital investment and savings, Keane and Wolpin (2001) and Van der Klaauw and Wolpin (2008). Neither of those papers, however, models the continuous choice of hours,

although they allow for several discrete alternatives.[102] The main focus of the Keane and Wolpin (2001) paper is on schooling choice (not labor supply), so we discuss it in a later section. But their paper is of interest here because it assumes a CRRA utility function in consumption, and so, like Imai-Keane, provides an estimate of the key preference parameter η, which governs income effects in labor supply and intertemporal substitution in consumption. Keane and Wolpin (2001) obtain $\eta \approx -0.50$, which implies weaker income effects, and less curvature in consumption (i.e., higher willingness to substitute intertemporally), than much of the prior literature. Keane and Wolpin (2001, p. 1078) argue that the reason is that their work accommodates liquidity constraints, and that failure to do so may have led to a downward bias in estimates of η in prior work.[103]

Imai and Keane (2004) estimate that $\eta = -0.74$. This implies a somewhat lower intertemporal elasticity of substitution in consumption than the Keane and Wolpin (2001) estimate of $\eta \approx -0.50$ (that is, $\frac{1}{\eta} = -1.35$ vs. -2.0). But their estimate of η still implies weaker income effects on labor supply, and a higher willingness to substitute consumption intertemporally, than much of the prior literature. Instead of liquidity constraints (as in Keane and Wolpin (2001)), the Imai and Keane (2004) model "explains" the fact that young workers do not borrow heavily against higher future earnings by assuming age effects in the marginal utility of consumption. Both models provide a good fit to asset data over the life-cycle. Finally, Keane (2009b) uses the Imai-Keane estimates of γ and η to calibrate a simple two-period equilibrium model. He finds that welfare costs of labor income taxation are much larger than more conventional values of γ and η would suggest.

In summary, although the literature that uses dynamic programming models to study life-cycle labor supply, asset accumulation and human capital investment for males is quite small, it has produced important results. Specifically, it finds that the intertemporal elasticities of substitution for both labor supply and consumption are quite a bit larger than implied by earlier work. This, in turn, implies that tax effects on labor supply for males may be larger than conventionally thought. Clearly more work is called for to investigate the robustness of these results to alternative model specifications and data sources.

4.2. Job search

Along with the dynamic labor force participation model, among the first applications of the DCDP approach was to the estimation of models of job search—the transition from unemployment to employment. The labor supply and job search literatures have,

[102] Van der Klaauw and Wolpin (2008) estimate a collective model of the joint labor supply decisions of a married couple nearing retirement They allow for savings and human capital accumulation, incorporating as well a detailed representation of US social security system.rules. As noted in the introduction, we do not review the DCDP retirement literature in this chapter.

[103] The estimate of η in Van der Klaauw and Wolpin (2008) is -0.6, which is also in line with other estimates from the retirement literature. They also include liquidity constraints.

however, addressed different questions and followed distinct paths. To understand why that has been the case, recall that in the labor force participation model workers with the same characteristics, and thus the same level of productivity, are offered the same wage. That is, in the wage offer function, $\log w_{it} = \log r + \gamma_0 + z_{it}\gamma + \eta_{it}$, r is assumed to be a market-level (for example, competitively determined) skill rental price and z_{it} and η_{it} are worker characteristics.

In contrast, the job search literature starts from the assumption that firms may offer different wages (skill rental prices, r) to identical workers within a given labor market. Then, the wage offer received by a worker of given characteristics from a firm j is $\log w_{itj} = \overline{\log r} + \gamma_0 + z_{it}\gamma + \zeta_j$, where $\overline{\log r}$ is the mean skill rental price in the labor market and where ζ_j reflects firm j's idiosyncratic component of the skill rental price. In the basic model, the accepted job lasts "forever" and individuals are not subject to productivity shocks. Given this wage structure, once an individual accepts a wage offer from a firm with a given ζ_j, their skill rental price is fixed for as long as they work for that firm.[104] The information set of the individual includes the distribution of ζ_j across firms, but not which firms are matched with particular values of ζ. Because there are more and less desirable firms, individuals have an incentive to engage in job search, that is, to look for a high wage firm. Job search is sequential. The difference in the labor force participation and job search models thus reflects the different assumptions made about the wage structure of the labor market.

The partial equilibrium search model has normally been used to understand a different phenomenon than has the labor supply model. Official labor force statistics distinguish among three mutually exclusive and exhaustive states, being employed, being unemployed and being out-of-the labor force. The distinction between the latter two states is based on whether an individual is actively seeking work. Both the labor force participation and search models consider only two states. In the labor force participation model, unemployment and out-of-the labor force are collapsed into one nonemployment state. In that model, it is assumed that a new wage offer is received every period and any individual will work at some offered wage. The search model conditions on individuals having already chosen unemployment over being out-of-the labor force and does not assume that a job offer necessarily arises each period. Because of this difference, labor force participation models have been applied to low frequency data based on the employment-nonemployment dichotomy, commonly at the annual level and often for women, while job search models have been applied to high frequency data, for example, at the weekly level, based on the employment-unemployment dichotomy.

The structural implementation of the standard partial equilibrium job search model was first considered by Wolpin (1987) and Van den Berg (1990), building upon a

[104] Wages would grow deterministically if z contains age or job tenure.

nonstructural literature that had begun a decade or more before.[105] The nonstructural empirical literature was focused on the evaluation of the effect of UI programs, more specifically, on the estimation of the impact of unemployment benefits on the duration of unemployment and wages. The empirical approach in that literature was (and is still) based, loosely and in some ways incorrectly, on the sequential job search model first formalized by McCall (1970) and Mortensen (1970). The structural empirical literature, following the DCDP paradigm, is based on the explicit solution and estimation of the sequential model.

There have been many extensions and modifications of the standard model in the structural empirical literature. Within the standard framework, Stern (1989), extending the original contribution of Stigler (1961) to a sequential framework, allowed for simultaneous search, that is, for the submission of multiple job applications in a period. Blau (1991) dropped the assumption of wealth maximization, allowing for job offers to include not only a wage offer but also an hours offer. Ferrall (1997) incorporated all major features of the Canadian UI program. Gemici (2007) considered the joint husband and wife search-migration decision in an intra-household bargaining framework. Paserman (2008), adopting a behavioral approach, allowed for hyperbolic discounting.

The standard model has also been extended beyond the consideration of the single transition from unemployment to employment. Wolpin (1992) incorporated job-to-job transitions and both involuntary and voluntary transitions into unemployment.[106] Rendon (2006) allowed for a savings decision in a setting where agents can also transit, both through quits and layoffs, from employment to unemployment.[107] Both of these latter papers also allowed for wage growth with the accumulation of work experience, employer-specific (tenure) in the case of Rendon and both general and employer-specific in the case of Wolpin.

As noted, the standard job search model assumes that *ex ante* identical workers may receive different wage offers, or analogously, that the same unemployed worker may receive different offers over time. Diamond (1970) showed that with the assumptions of the standard job search model, in a game in which firms are aware of worker search strategies, the wage offer distribution will be degenerate at the worker's reservation wage or outside option. This result led to the development of models in which wage dispersion could be rationalized as an equilibrium outcome, which in turn led to a structural

105 A number of these earlier papers appeared in a 1977 symposium volume of the *Industrial and Labor Relations Review*. Most relevant in that volume are papers by Classen, and the comment on them by Welch (1977).

106 Burdett (1978) extended the standard unemployment search model to allow for search on the job.

107 It is a common theme in the structural literature to build upon and extend the theoretical literature in developing estimable models. This is the case with Rendon's (2006) paper, which builds on the earlier work of Danforth (1979). Lentz (2009) also structurally estimates a sequential search model with savings. Unlike the standard model, the wage offer distribution is taken to be degenerate and agents choose their search intensity, which affects the rate at which job offers are received. Lentz uses the model to empirically determine the optimal unemployment insurance scheme.

empirical literature. The empirical literature has focused on two kinds of models, those based on a search-matching-bargaining approach to wage determination (Diamond and Maskin, 1979; Mortensen, 1982; Wolinsky, 1987) and those based on wage-posting by firms that gain monopsony power through search frictions (Albrecht and Axell, 1984; Burdett and Mortensen, 1998). These models, not only rationalized wage dispersion, but also allowed for quantification and policy analyses in an equilibrium setting, for example, changes in UI benefits or changes in the level of the minimum wage.

The empirical implementation of equilibrium search models has become a major strand of the structural job search literature.[108] Embedded within those models are different variants of the standard partial equilibrium search model and, in that sense, the development of the DCDP estimation approach was a critical precursor. However, given that the modeling has gone well beyond the partial equilibrium search model to which the DCDP approach has direct application, it would take us too far afield to provide a review of that literature. For such a review, we would refer the reader to the chapter by Mortensen and Pissarides in the Handbook of Labor Economics (Volume 3b, 1999) or the more recent survey by Eckstein and Vandenberg (2007).

In the rest of this section we review the structural empirical literature on the partial equilibrium job search model. Because the structural literature is explicitly connected to the theory, we first present the formal structure of the standard job search model and show how the nonstructural empirical literature can be interpreted in the context of the job search model. We then discuss conditions for identification and methods of estimation. Finally, we describe three empirical papers that have estimated extended versions of the standard model, and thus exemplify the nature of scientific progress in the structural literature, as we discussed in the introduction of this chapter, and we report empirical findings from counterfactual experiments in those papers.

4.2.1. The standard discrete-time job search model

In the discrete time formulation, an unemployed individual receives a job offer in each period with probability q. Wage offers are drawn from the known cumulative distribution function, $F(w)$. An accepted job offer (and its concomitant wage) is permanent. While unemployed an individual receives b, unemployment benefits (if eligible) net of the cost of search. The individual is assumed to maximize the present discounted value of net income. We consider both infinite and finite horizon models, which have somewhat different empirical implications and implications for the identification of model parameters.[109]

[108] Examples of papers based on wage posting models include Eckstein and Wolpin (1990), Kiefer and Neumann (1993), Van den Berg and Ridder (1998), Bontemps et al. (1999, 2000) and Postel-Vinay and Robin (2002). Those based on search-matching-bargaining.models include, among others, Eckstein and Wolpin (1995), Cahuc et al. (2006) and Flinn (2006).

[109] See Mortensen (1986) for the continuous time case.

Infinite horizon model

The value of a wage offer of w, given a discount factor of $\delta(=\frac{1}{1+r})$, is

$$\begin{aligned} W(w) &= w + \delta w + \delta^2 w + \cdots \\ &= w + \delta W(w) \text{ or} \qquad (133) \\ W(w) &= \frac{w}{1-\delta}. \qquad (134) \end{aligned}$$

In any period, the value of continuing to search, V, either because an offer was rejected or an offer was not received, consists of the current period payoff, b, plus the discounted expected value of waiting another period. In that case, if an offer is received, with probability q, the individual chooses between the maximum of the value of working at a wage w, $W(w)$, or continuing to search and receive V. If no offer is received, which occurs with probability $1-q$, the individual must continue to search and receives V. Thus the alternative-specific value function, the Bellman equation, for the search choice is

$$V = b + \delta[qE\max(W(w), V) + (1-q)V]. \qquad (135)$$

Rearranging, yields

$$V(1-\delta) = b + \delta qE\max(W(w) - V, 0), \qquad (136)$$

which has a unique solution for $V > 0$, as long as the cost of search is not so large as to make the right hand side negative.[110] Defining w^*, the reservation wage, to be the wage offer that equates the value of search and the value of accepting the job, that is,

$$w^* = (1-\delta)V, \qquad (137)$$

with a little further algebra, we obtain the following implicit equation for the reservation wage (which must have a unique solution given that V does):

$$w^* = b + \frac{q}{r}\int_{w^*}^{\infty}(w - w^*)\mathrm{d}F(w).^{111} \qquad (138)$$

[110] The LHS is linearly increasing in V and passes through the origin. The RHS is monotonically decreasing in V until it reaches b, and is then constant. There will be a unique intersection, and a unique $V > 0$, as long as $b > -\beta qE\max(W(w), 0)$.

[111] In a continuous time model in which the arrival of offers follows a Poisson process with parameter λ, the implicit reservation wage equation is identical except that the instantaneous arrival rate (λ) replaces the offer probability, q.

Thus, the reservation wage is a function of b, $\frac{q}{r}$ and $F(w)$:

$$w^* = w^*\left(b, \frac{q}{r}, F\right).^{112} \tag{139}$$

The individual accepts any wage offer that exceeds the reservation wage and declines offers otherwise.

Although the reservation wage is a deterministic function, the length of an unemployment spell is stochastic because the timing and level of wage offers are probabilistic. Thus, measures of the outcomes of search, such as the duration of unemployment spells and the level of accepted wages, are probabilistic. In particular, the survivor function, the probability that the duration of unemployment is at least as large as some given length, is

$$\Pr(T_u \geq t_u) = [qF(w^*) + (1-q)]^{t_u} \tag{140}$$

$$= [1 - q(1 - F(w^*))]^{t_u}. \tag{141}$$

The term inside the brackets in (140) is the probability of receiving an offer in a period and rejecting it (because it is below the reservation wage) plus the probability of not receiving an offer. The cdf, pdf and hazard function are:

$$\text{cdf} : \Pr(T_u < t_u) = 1 - [1 - q(1 - F(w^*))]^{t_u}, \tag{142}$$

$$\text{pdf} : \Pr(T_u = t_u) = [1 - q(1 - F(w^*))]^{t_u} q(1 - F(w^*)) \tag{143}$$

$$\text{Hazard Function} : \Pr(T_u = t_u | T_u \geq t_u) = \frac{\Pr(T_u = t_u)}{\Pr(T_u \geq t_u)} = q(1 - F(w^*)) = h. \tag{144}$$

As seen, the survivor function, the cdf and the pdf can all be written as functions of the hazard rate, the exit rate from unemployment conditional on not having previously exited. From (144), it can be seen that the hazard rate is constant. Thus, in a homogeneous population, the infinite horizon search model implies the absence of duration dependence.

Given parameter values, mean duration is given by

$$E(t_u) = \sum_0^\infty t_u \Pr(T_u = t_u)$$

$$= q(1 - F(w^*))^{-1} = \frac{1}{h}. \tag{145}$$

[112] Given a distributional assumption for $F(w)$, the solution for the reservation wage involves numerically (if, as is for most distributions the case, there is no closed form solution) solving a nonlinear equation. An alternative solution method would be to start from the reservation wage for the final period of a finite horizon problem (see below) and iterate on the reservation wage until it converges. Convergence is assured because the value function is a contraction mapping (see Sargent (1987)).

Notice that mean duration is simply one over the hazard rate.[113] Likewise, the mean of the accepted wage is

$$E(w|w \geq w^*) = \int_{w^*}^{\infty} \frac{w}{1 - F(w^*)} \mathrm{d}F(w), \tag{146}$$

which clearly is larger than the mean of the wage offer distribution.[114]

In addition to implying a constant hazard rate, the infinite horizon model has predictions about the impact of changes in b, $\frac{q}{r}$ and $F(w)$ on the reservation wage, on the hazard rate and on the moments of the accepted wage distribution. It thus is, in principle, possible to test the theory. Comparative static effects of the hazard rate (and thus mean duration) with respect to its arguments are (see Mortensen (1986)):

$$\frac{\mathrm{d}h}{\mathrm{d}b} = -\lambda f(w^*) \left(\frac{1}{1 + h/r} \right) < 0, \tag{147}$$

$$\frac{\mathrm{d}h}{\mathrm{d}q} = -f(w^*) \left(\frac{w^* - b}{1 + h/r} \right) + (1 - F\left(w^*\right)) \gtrless 0, \tag{148}$$

$$\frac{\mathrm{d}h}{\mathrm{d}\mu} = -qf(w^*) \left(\frac{h/r}{1 + h/r} \right) > 0, \tag{149}$$

$$\frac{\mathrm{d}h}{\mathrm{d}s} = -qf(w^*) \left(\frac{(\lambda/r) \int_0^{w^*} F_s(w, s) \mathrm{d}w}{1 + h/r} \right) < 0. \tag{150}$$

An increase in the level of unemployment compensation benefits increases the reservation wage and reduces the unemployment hazard rate (147). An increase in the offer probability has an ambiguous effect; it increases the reservation wage, which reduces the hazard rate, but also directly increases the hazard rate through the higher offer probability (148). It turns out that for a certain class of distributions (log concave), the latter effect dominates (Burdett and Ondrich, 1985). Increasing the mean of the wage offer distribution (149), μ, increases the hazard rate; although an increase in μ increases the reservation wage, the increase is less than one for one. Finally, increasing the mean preserving spread of the distribution (150), s, reduces the hazard because an increase in the mass of the right tail of the wage offer distribution increases the payoff to search, thus increasing the reservation wage. An additional set of implications follow about the mean of the accepted wage; anything that increases the reservation wage also increases the mean accepted wage.

Finite horizon model

Spells of unemployment tend to be short (weeks or months) in relation to an individual's life span. A finite lifetime would not seem, therefore, to be a reason to explore the

[113] Wolpin (1995) provides a proof.

[114] The distribution of accepted wages is the truncated distribution of wage offers, namely, $f(w|w \geq w^*) = \frac{f(w)}{1-F(w^*)}$.

implications of a finite horizon search model. On the other hand, it is reasonable to assume that individuals generally will not be able to self-finance extended periods of unemployment and that external borrowing is limited. One can think, then, of the finite horizon as corresponding to the maximal unemployment period that can be financed through internal and external funds, although we continue to assume that once a job is accepted it lasts forever, that is, the horizon is infinite subsequent to accepting a job. In addition to the previous notation, we denote by T the end of the search horizon. To close the model, it is necessary to specify the value function if the terminal period is reached without having accepted a job. We assume that in that case the individual receives b forever.[115]

Without going into the details, the reservation wage path can be shown to satisfy the following difference equation:

$$\frac{w^*(t)}{1-\delta} = b + \frac{\delta}{1-\delta} w^*(t+1) + \frac{\delta}{1-\delta} q \int_{w^*(t+1)}^{\infty} (w - w^*(t+1)) \mathrm{d}F(w) \quad \text{for } t < T, \tag{151}$$

$$w^*(T) = b. \tag{152}$$

Notice that (151) reduces to the implicit reservation wage equation for the infinite horizon problem if $w^*(t) = w^*(t+1)$. Given a distributional assumption for wage offers, $F(w)$, the solution of the finite horizon reservation wage path can be obtained numerically by starting from period T and working backwards.[116]

In the finite horizon case, the reservation wage is decreasing in the duration of the spell, $\frac{\mathrm{d}w^*(t)}{\mathrm{d}t} < 0$, rather than being constant as in the infinite horizon case. In addition, the reservation wage is bounded from below by b (at T), and from above by the infinite horizon reservation wage (w^*). The hazard rate is thus increasing in t, $\frac{\mathrm{d}h(t)}{\mathrm{d}t} = -\lambda f(w^*) \frac{\mathrm{d}w^*(t)}{\mathrm{d}t} > 0$. Thus, the longer the spell duration, the greater the exit rate. The important property of the finite horizon reservation wage is that it depends on the time left until the horizon is reached. The reservation wages are equal under two different horizons not when they have the same amount of time elapsed since beginning the spell of unemployment, but when they have the same amount of time left until the horizon is reached ($T - t$).

Nonstructural (parametric) approach to estimation

The early nonstructural approach to estimating the job search model was regression based. The primary concern of that literature, as well as the later literature based on hazard modeling, was the estimation of the impact of unemployment benefits on the

115 An alternative would be to assume that once the terminal period is reached, the individual accepts the next offer that arrives, in which case the reservation wage at T is zero.

116 See Wolpin (1987), for the particular case in which $F(w)$ is either normal or log normal.

duration of unemployment and post-unemployment wages.[117] The regression (or hazard rate) specification in the nonstructural approach was motivated by the standard job search model. Classen (1977) provides a clear statement of the connection between the theory and the regression specification. The latter is given by

$$D = \alpha_0 + \alpha_1 WBA + \sum \alpha_i X_i + u_D, \tag{153}$$

$$Y = \beta_0 + \beta_1 WBA + \sum \beta_i X_i + u_Y \tag{154}$$

where D is spell duration, Y is a measure of the post-unemployment wage, WBA is the weekly UI benefit amount and the X's are "proxies" for a worker's skill level, the cost of search and the job offer rate. As Classen notes, the determinants of both spell duration and the post-unemployment wage should be exactly the same as they are both optimal outcomes derived from the search model. Among the proxy variables used in Classen's analysis are demographics, such as age, race and sex, and a measure of the wage on the job held prior to beginning the unemployment spell. Although not included in the Classen study, other variables often included in this type of specification are education, marital status, number of dependents and a measure of aggregate unemployment in the relevant labor market.

A test of the theory amounts to a test that benefits increase both expected duration and the mean accepted wage, that is, that α_1 and β_1 are both positive. Any further test of the theory would involve specifying how proxy variables are related to the structural parameters, q, the offer probability, and $F(w)$, the wage offer distribution.

Classen is clear as to the purpose of including the pre-unemployment wage, namely as a proxy, most directly perhaps for the mean of the wage offer distribution. However, although the inclusion of that variable or, as is also common, of the replacement rate, the ratio of the benefit level to the pre-unemployment wage, was and continues to be standard in the nonstructural literature, the need for stating a rationale has been lost. As has been pointed out elsewhere (Wolpin, 1995), the inclusion of the pre-unemployment wage (or the replacement rate) cannot be justified by the standard search model, that is, given perfect measures of q and $F(w)$, it would not have any impact on search outcomes. More importantly, however, given its ubiquitous use, is that its inclusion leads to "proxy variable bias," as explained below.

Of course, the rationale for its inclusion is to avoid omitted variable bias. For example, suppose that the pre-unemployment wage is meant to proxy the mean of the wage offer distribution faced by individual i, μ_i. Now, UI benefits are usually tied to the pre-unemployment wage, at least up to some limit. Thus, if some determinants of μ_i

[117] There are statistical issues better handled by specifying the hazard function, such as dealing with incomplete spells and time-varying regressors. See Meyer (1990) for an example of this approach. The issues we raise, however, are easier to demonstrate in a regression framework, but hold in the hazard framework as well.

are omitted (we do not have a perfect measure of $F(w)$), variation in the benefit level will reflect, in part, the fact that those with higher pre-unemployment wages have higher μ_i's. In that case, a positive correlation between UI benefit levels and the pre-unemployment wage will lead to a negative bias in the effect of UI benefits on the duration of unemployment (recall that the higher is μ_i, the greater the hazard rate).

Although omitted variable bias is well understood, the potential for bias introduced by using proxy variables is less well appreciated. The source of the problem is that, in the context of a search model, the pre-unemployment wage must have been the outcome of a search during a prior unemployment spell. To isolate the effect of using this proxy, assume that the duration of the prior unemployment spell was governed by the same behavioral process and fundamentals as the current spell. Such an assumption is consistent with a model in which there are exogenous layoffs and unemployment spells are renewal processes (the stochastic properties of all unemployment spells are the same).[118] In particular, suppose that only the benefit level and the mean of the wage offer distribution vary (say, geographically) in the sample. Then, taking deviations from means (without renaming the variables, to conserve on notation), the duration equation is

$$D_i = \pi_{11} b_i + \pi_{21} \mu_i + v_{1i}. \tag{155}$$

Assuming μ_i is unobserved, and thus omitted from the regression, the bias in the OLS estimator of π_1 is given by

$$E(\widehat{\pi}_{11} - \pi_{11}) = \pi_{21} \frac{\sigma_\mu^2}{\sigma_b^2} \beta_{b,\mu}, \tag{156}$$

where $\beta_{b,\mu}$ is the regression coefficient of b on μ. Thus, if b and μ are positively correlated and $\pi_{21} < 0$ as theory suggests, the bias in the estimated effect of UI benefits on duration will be negative.

Now, because the pre-unemployment wage, Y_{-1}, is the result of a prior search, it will have the same arguments as (155), namely

$$Y_{i,-1} = \pi_{12} b_i + \pi_{22} \mu_i + v_{2i} \tag{157}$$

$$= \pi_{22} \mu_i + \omega_i, \tag{158}$$

where $\omega_i = \pi_{12} b_i + v_{2i}$ and where $E(\omega b_i) \neq 0$ given the definition of ω_i. To derive a regression equation that includes b_i and $Y_{i,-1}$, solve for μ_i in (158) and substitute

[118] The assumption that new unemployment spells are renewal processes rules out any structural connection between spells; for example, it rules out that the benefit level depends on the pre-unemployment wage.

into (155), yielding

$$D_i = \pi_{11} b_i + \pi_{21} \frac{(Y_{i,-1} - \omega_i)}{\pi_{22}} + v_{1i} \tag{159}$$

$$= \pi_{11} b_i + \frac{\pi_{21}}{\pi_{22}} Y_{i,-1} + v'_{1i}, \tag{160}$$

where $v'_{1i} = -\frac{\pi_{21}}{\pi_{22}}\omega_i + v_{1i}$. We are interested in whether the OLS estimator for π_{11} is biased, that is, whether $E(b_i v'_{1i} | w^a_{i,-1}) = E(-\frac{\pi_{21}\pi_{12}}{\pi_{22}} b_i^2 | \pi_{22}\mu_i + \pi_{12} b_i + v_{2i} = w^a_{i,-1}) = 0$. It is easiest to see whether this holds by explicitly deriving the bias expression. It is given by

$$E(\widehat{\widehat{\pi}}_{11} - \pi_{11}) = \pi_{21}^2 \frac{\sigma_\omega^2 \sigma_\mu^2}{\sigma_b^2 \sigma^2_{w^a_{-1}} - (\sigma_{b w^a_{-1}})^2} (\pi_{21}^{-1} \beta_{b,\mu} - \pi_{22} \beta_{b,\omega}). \tag{161}$$

The bias is zero if either $\sigma_\omega^2 = \pi_{12}^2 \sigma_b^2 + \sigma_{v_2}^2 = 0$ or if $\pi_{21}^{-1} \beta_{b,\mu} - \pi_{22} \beta_{b,\omega} = 0$, either of which would be a fortuitous property of the sample.

What is the relationship between the biases from omitting μ_i (omitted variable bias) versus including $w^a_{i,-1}$ (proxy variable bias)? It turns out that a sufficient condition for the bias from omitting the pre-unemployment wage to be smaller than from including it is that $\beta_{b,\omega} = \pi_{12} \sigma_b^2 = 0$, which only will hold if $\pi_{12} = 0$, a violation of the theory. In general, the biases cannot be ordered and it is unclear which is the better strategy to follow to minimize the bias. The implicit (that is, without justification) assumption made by almost all researchers is that omitted variable bias is greater than proxy variable bias in this context. Moreover, if the benefit level varies with the pre-unemployment wage and we have good measures of the mean of the wage offer distribution, variation in the benefit level from this source is helpful in identifying the UI benefit effect. The variation in the pre-unemployment wage around the mean of the offer distribution that induces benefit variation is purely due to random draws from the wage offer distribution.

Researchers adopting the nonstructural approach have universally included the pre-unemployment wage, and thus, assumed that omitted variable bias is greater than proxy variable bias. There is a larger point reflected by this choice, namely the importance of theory in empirical work. Structural work requires that all variables be explicitly accounted for in the model. A similar standard for nonstructural work might have revealed the existence of the choice between omitted variable and proxy variable bias, a choice not explicitly acknowledged in the nonstructural literature.[119]

[119] As we noted, it is also usual to include some aggregate labor market statistic like the local unemployment rate. The idea is that the aggregate statistic reflects labor market demand and so will affect the offer rate or the wage offer distribution. As shown in Wolpin (1995), because the aggregate statistic is simply the aggregation of the search decisions over the unemployed population, it does not reflect solely demand conditions and estimates of UI benefit effects suffer from proxy variable bias.

Structural approach to estimation

Identification: Identification is no less an important issue in structural empirical work than in nonstructural work. We consider identification of the standard search model parameters assuming we have data from a homogeneous population on durations of unemployment and on accepted wages.[120]

To establish identification, it is useful to rewrite the reservation wage implicit Eq. (138) as

$$w^* = b + \frac{q}{r}\int_{w^*}^{\infty}(w - w^*)\mathrm{d}F(x) \tag{162}$$

$$= b + \frac{q}{r}[E(w|w \geq w^*)(1 - F(w^*)) - w^*(1 - F(w^*))] \tag{163}$$

$$= b + \frac{h}{r}[E(w|w \geq w^*) - w^*], \tag{164}$$

where the last equality uses (144). From the accepted wage data, note that a consistent estimator of the reservation wage is the lowest observed wage: $\lim_{n\to\infty}\Pr(|w_{\min} - w^*|) = 0$.[121] Then, given an estimate of h from the duration data, and recognizing that from the accepted wage data, we can also obtain an estimate of $E(w|w \geq w^*)$, we can identify b (which includes the unobserved cost of search) if we take r as given. This result does not require a distributional assumption for wage offers.

We cannot, however, separate q and F without a distributional assumption. Although we know $F(w|w \geq w^*)$, given an estimate of w^*, we can recover the wage offer distribution, $F(w)$, only if it is possible to recover the untruncated distribution from the truncated distribution. Obviously, that cannot be done without making a distributional assumption. Assuming that $F(w)$ is recoverable from the accepted wage distribution, then from $h = q(1 - F(w^*))$, we can recover the offer probability, q.

Recoverability of the wage offer distribution is not always possible. It is useful for later reference to consider an example taken from Flinn and Heckman (1982). Assume that wage offer distribution is Pareto, that is, having pdf

$$f(w) = \varphi w^{\gamma}, \quad \varphi = \frac{-(\gamma + 1)}{c^{\gamma+1}} \quad c \leq w \leq \infty. \tag{165}$$

Notice that the support of the distribution is bounded from below by a constant c. The density of accepted wages is

120 This section follows the development in Flinn and Heckman (1982) and Wolpin (1995).

121 The minimum observed wage is a superconsistent estimator of the reservation wage in that it converges at rate N. This leads to nonstandard asymptotics in the likelihood estimation of the search model (see Flinn and Heckman (1982) and Christensen and Kiefer (1991)).

$$f(w|w \geq w^*) = \frac{\varphi w^\gamma}{1 - F(w^*)} \tag{166}$$

$$= \frac{-(\gamma+1)}{(w^*)^{\gamma+1}} w^\gamma, \quad w \geq w^* \geq c. \tag{167}$$

Given an estimate of w^*, from the minimum observed wage, and of the conditional density, we can recover γ. However, there are many values of c that are consistent with the truncated distribution. We, thus, cannot identify φ, which means we cannot identify the wage offer distribution, F. As already noted, the consequence of this lack of identification is that we cannot separate q and F, or, more specifically, φ. To see that explicitly, write the reservation wage equation and the hazard function under the Pareto distribution,

$$w^* = b + \frac{q}{r} \int_{w^*}^{\infty} (w - w^*) \mathrm{d}F(x)$$

$$= b + \frac{q\varphi}{r} \left(\frac{(w^*)^{\gamma+2}}{r(\gamma+1)(\gamma+2)} \right); \tag{168}$$

$$h = \frac{-\lambda\varphi}{\gamma+1} (w^*)^{\gamma+1}. \tag{169}$$

Because q and φ only enter multiplicatively as $q\varphi$, it is impossible to separately identify them. Fortunately, most of the commonly used distributions for wage offer functions, for example, the lognormal distribution, are recoverable from the distribution of accepted wages. However, there is an important lesson to draw, namely that parametric assumptions do not always assure identification.

The analysis of identification in the finite horizon case is similar. The reservation wage at each period can be consistently estimated from the period-specific minimum observed wages. Analogous to (164), we can write the implicit reservation wage equation as

$$w_t^* = b(1-\delta) + \delta w_{t+1}^* + \delta h_{t+1}[E(w|w \geq w_{t+1}^*) - w_{t+1}^*].^{122} \tag{170}$$

This equation must hold exactly at each time t. As long as there are durations of unemployment that extend through more than two periods, that is, given $w_t^*, w_{t+1}^*, w_{t+2}^*, h_{t+1}, h_{t+2}, E(w|w \geq w_{t+1}^*)$ and $E(w|w \geq w_{t+2}^*)$, δ (or r) and b can be separately identified by solving the two difference equations for the two unknowns, δ and b. Recall that this separation was impossible in the infinite horizon case. Moreover, if we have more than three periods of data, the model is rejectable. Identifiability of q and F, however, still requires recoverability of the wage offer distribution.

Likelihood function: The likelihood function for the search model takes an analogous form as that for the binary labor force participation model, given data on unemployment

[122] Setting $w_t^* = w_{t+1}^* = w^*$ and $h_{t+1} = h$ yields the infinite horizon implicit reservation wage function.

durations and accepted wages. Consider the likelihood contribution of an individual solving an infinite horizon search model, who has a completed unemployment spell of length d_i and accepted wage w_i^a:

$$L_i = \left[q \Pr(w < w^*) + (1-q)\right]^{d_i - 1} \times q \Pr(w_i^a > w^*, w_i^a) \tag{171}$$

$$= \left[q \Pr(w < w^*) + (1-q)\right]^{d_i - 1} \times q \Pr(w_i^a > w^* | w_i^a) f(w_i^a). \tag{172}$$

Note that the reservation wage, the solution to the implicit equation (138), is a function of the model parameters b, δ, q and $F(w)$. The first (bracketed) term in the likelihood is the probability that in each of the periods up to $d_i - 1$ the individual received an offer and rejected it or did not receive an offer. The second term is the probability that the individual received an offer of w_i^a in period d_i and accepted it. Individuals who have incomplete unemployment spells would contribute only the first term to the likelihood.

In the labor force participation model of Section 3.1.1, the probability of accepting an offer to work depended on both a random wage draw and a random taste draw. Thus, whether or not an individual worked conditional on a wage draw was probabilistic, because it depended on the taste draw that we do not observe. In the search model, however, for any given value of the reservation wage (or the parameters that determine it), whether or not an individual works conditional on a wage draw is deterministic, that is, its probability, $\Pr(w_i^a > w^* | w_i^a)$ in (172), is either one or zero. In order that the likelihood not be degenerate, the reservation wage must be less than the lowest accepted wage in the sample; for this reason, the minimum observed wage in the sample is the maximum likelihood estimate of the reservation wage. That would not create an issue if observed wages were all reasonable. However, in most survey data sets the lowest reported wage is often quite small, less than one dollar or even only a few cents. Such outliers would potentially have an extreme effect on the estimates of the structural parameters. One remedy would be to be to trim the wage data, say by whatever percent led to a "reasonable" lowest wage. However, that would be essentially choosing the reservation wage by fiat. A second alternative would be to add another error to the model, for example, by allowing the cost of search to be stochastic, in which case the reservation wage would be stochastic.[123] A very low accepted wage would be consistent with the individual having drawn a very high search cost.

Of course, adding another source of error does not deal with what is the likely root cause of the problem, which is that wages are not accurately reported.[124] That "fact" has

[123] The implicit reservation wage Eq. (138) would no longer hold in this case. In particular, the integration would have be taken also over the distribution of the unobserved cost of search, recognizing that the reservation wage would be a function of that cost.

[124] In some instances, wage rates that are directly reported in hours or weeks are inaccurate. In other cases, wage rates are derived from a division of earnings, reported over a longer period, say annually, and hours worked reported over that period. The inaccuracy arises from a seeming mismatch in the time period between earnings and hours.

led researchers to directly allow for measurement error in the reported wage. Introducing measurement error not only accounts for a real feature of the data, it is also convenient in that it does not require any change in the solution of the search problem. The reservation wage is itself unaffected by the existence of measurement error; the implicit reservation wage equation is still given by (138). Taking into account the existence of measurement error, letting w_i^{aR} be the reported accepted wage and w_i^{aT} the true accepted wage, the likelihood function (172) becomes

$$\begin{aligned} L_i &= \left[q \Pr(w_i^T < w^*) + (1-q)\right]^{d_i-1} \times q \int_{w_i^{aT}} \Pr(w_i^{aT} > w^*, w_i^{aR}) \\ &= \left[q \Pr(w_i^T < w^*) + (1-q)\right]^{d_i-1} \\ &\quad \times q \int_{w_i^{aT}} \Pr(w_i^T > w^* | w_i^{aR}, w_i^{aT}) g(w_i^{aR} | w_i^{aT}) \mathrm{d}F(w^{aT}) \\ &= \left[q \Pr(w_i^T < w^*) + (1-q)\right]^{d_i-1} \\ &\quad \times q \int_{w_i^{aT}} \Pr(w_i^{aT} > w^* | w_i^{aT}) g(w_i^{aR} | w_i^{aT}) \mathrm{d}F(w^{aT}), \end{aligned} \tag{173}$$

where $g(w_i^{aR}|w_i^{aT})$ is the distribution of the measurement error and where the third equality emphasizes the fact that it is the true wage only and not the reported wage that affects the acceptance probability. The most common assumption in the literature is that the measurement error is multiplicative, that is, proportional to the true wage.

The estimation of the finite horizon search problem when there are extreme low-wage outliers is particularly problematic. Recall that the reservation wage is declining with duration. Thus, if an outlier observation occurs at an early duration, the entire subsequent path of reservation wages must lie below the reservation wage at that early duration. The incorporation of measurement error is, therefore, critical for estimation. The analogous likelihood contribution for an individual for the finite horizon model, which takes into account that the reservation wage is duration dependent, is

$$L_i = \prod_{j=1}^{d_i-1} \left[q \Pr(w_j^T < w_j^*) + (1-q_j)\right] \times q \Pr(w_{d_i}^{aT} > w_d^*, w_d^{aR}). \tag{174}$$

The estimation of the partial equilibrium search model involves the same iterative process as for the labor force participation model, namely numerically solving a dynamic programming problem at trial parameters and maximizing the likelihood function. The generality of the DCDP approach has allowed researchers considerable flexibility in modeling choices. Thus, researchers have adopted different distributional assumptions

and have extended the standard search model in a number of directions that we have already mentioned.

As has been generally true in the DCDP literature, the theoretical models that serve as their foundation cannot directly be taken to the data. In the case of the standard search model, neither the infinite horizon nor the finite horizon model can fit the generally observed fact that the hazard rate out of unemployment declines with duration. Recall that the hazard rate is constant in the infinite horizon case and increasing in the finite horizon case. There are several ways to deal with this mismatch between the data and the models. In the infinite horizon model, introducing unobserved heterogeneity in model fundamentals, such as the cost of search, the offer probability and/or the wage offer distribution, can produce negative duration dependence in the population hazard while maintaining stationarity at the individual level. In the finite horizon case, allowing for time dependencies in model fundamentals, such as allowing offer probabilities to decline with duration, can create negative duration dependence, in this case at the individual level as well as at the population level.

Selected literature

Rather than do a comprehensive review of the contributions of DCDP modeling to the empirical (partial equilibrium) search literature, we illustrate the broad range of model specifications and the usefulness of the approach for policy evaluation with three examples.

Rendon (2006): In this first example, Rendon (2006) extends the standard finite horizon search model to allow for exogenous job loss (layoffs), for on-the-job search and for savings in the presence of borrowing constraints. Recall that the standard finite horizon search model imposes a terminal search period, with the putative rationale being that the individual cannot search indefinitely due to a limit on borrowing. However, because this limitation is not an explicit part of the model, the terminal value function (or equivalently, the terminal reservation wage) is not determined as part of the model. Rendon (2006) fills this lacuna in the structural empirical literature.[125]

Rendon, building on theoretical papers by Danforth (1979) and others and on the DCDP model of Wolpin (1992), considers a job search model with the following features:

1. Individuals maximize the present discounted value of lifetime utility. Flow utility is a CRRA function in consumption. Individuals are finitely lived and exogenously retire at a known time prior to the end of life. Time is discrete. Prior to retirement, the individual is in one of two employment states, unemployed or employed.

125 Theoretical models in which job searchers faced borrowing constraints appeared much earlier, starting with Danforth (1979). However, formal empirical implementation did not become feasible until the development of estimation methods for DCDP models.

2. As in the standard search model, in each period of unemployment, the individual receives a job (wage) offer with a positive probability. If an offer is received, the individual makes an acceptance-rejection decision. The individual enters each period with some level of assets (positive or negative) and decides the level of assets to carry forward to the next period. Income while unemployed consists of the interest return (or payment, if assets are negative) on assets, and unemployment compensation benefits plus other family transfers minus the cost of search.
3. In each employment period, the individual faces a positive probability of layoff as well as a positive probability of receiving a wage offer from another employer. Regardless of whether an offer is received, the individual can decide to quit and become unemployed. As in the unemployment state, the individual decides on the level of assets to carry forward into the next period. Income while employed consists of the interest return (or payment, if assets are negative) on current assets plus the wage. Wages grow, starting at the initial accepted wage, deterministically with job tenure.
4. The level of assets that an individual holds in any period can be negative; the individual can carry debt, but the amount of debt cannot fall below the present value of the amount the individual can pay back with certainty (the Hakansson-Miller or "natural" borrowing limit). Because the individual can with some small probability be unemployed until the retirement date, the only certain income each period is the amount of unemployment income given by UI benefits and family transfers net of search costs.

In this model, individuals generally accumulate assets while employed as insurance against future unemployment spells and decumulate assets to finance search while unemployed. An individual's reservation wage, as in the standard model, declines with duration as assets are run down. Individuals who start an unemployment spell (at the same life cycle point, say, due to a layoff) with higher assets (having randomly drawn a higher acceptable wage offer on the prior spell of unemployment) have higher reservation wages, longer unemployment spells and higher accepted wages. Thus, wages will not only be positively correlated across employment spells generated by job-to-job transitions, but also across unemployment spells separated by layoffs.

Unemployment spells are not only generated by exogenous layoffs, but also by voluntary quits into unemployment, even though wages on the job are non-stochastic. This behavior can arise when the offer probability is greater while unemployed than while employed. Consider an unemployed individual who, having not received frequent offers or only received offers at low wages, has drawn down assets to finance consumption, perhaps even hitting the borrowing constraint. The individual in this situation optimally accepts a low wage job. The individual, once employed, will begin to accumulate assets as insurance. At some time, the individual, having not received any higher wage offer from another firm will have accumulated sufficient assets for it to be

optimal to quit into unemployment, financing consumption with the accumulated assets, to take advantage of the higher job offer rate during unemployment.[126]

As seen by this discussion, the existence of voluntary quits requires a certain parameter configuration. Given data in which voluntary quits arise, this parameter configuration must be an outcome of the estimation. Most DCDP models, like this search model, have the characteristic that model predictions are parameter dependent. This characteristic does not imply that these models do not have rejectable restrictions. Indeed, DCDP models are generally highly restrictive. Recall that the standard finite horizon model was rejectable, not in the conventional way of testing comparative static predictions, but because only a few parameters determined the entire profile of reservation wages. Tests of DCDP models are best thought of as tests arising from cross-equation restrictions. Models like the one estimated in Rendon also have such cross-equation restrictions, but they do not have easily derived analytical forms. However, to the extent that those cross-equation restrictions are seriously violated, the model will not be able to fit the data very well. Tests of model fit are (imperfect) tests of the model's implicit cross equation restrictions.

In addition to the discrete state variable, job tenure, the search problem in Rendon has two continuous state variables, assets and the accepted wage. The latter is a state variable because the accepted wage is permanent (and wage growth is deterministic), which implies that reservation wage for accepting an offer from another employer depends on the wage at the current employer. Rendon solves the DP model by discretizing assets and wages, a method we discussed above. However, given the fine discretization he used, it was not tractable to solve and estimate the model over his postulated 40.5-year working life on a quarterly basis, that is for 162 quarters. To make it tractable, Rendon solved the model on a quarterly basis for the first 12.5 years, then on an annual basis for the next 8 years and finally a biannual basis for the next 20 years.[127]

Rendon's model is extremely parsimonious. It contains only 12 parameters that must account for all of the labor force transitions, wages and assets observations of a sample of white male high school graduates over the first 40 quarterly subsequent to their graduation.[128] The model is estimated by maximum likelihood. As is conventional in the DCDP literature, Rendon computes chi-square statistics for the match between the actual data and data generated by the model estimates for a wide range of statistics. The quality of the fit of the model is mixed.

The estimated model is used to perform a number of counterfactual exercises. In particular, Rendon considers the impact on labor market outcomes of relaxing the extent to which borrowing constraints are binding. In the quantitative experiment he

126 The search model does not have to incorporate savings for it to be optimal for individuals to quit into unemployment. It is sufficient that there be a finite horizon (retirement) and either that the offer rate be higher in unemployment or that there be a wage return to general experience (Wolpin, 1992).

127 This procedure for reducing the size of the state space follows Wolpin (1992).

128 Assets are only observed annually. The data are from the NLSY79.

performs, he finds that allowing agents to borrow up to 50 percent of the natural borrowing limit, as opposed to the estimated 10 percent in the baseline, would increase the duration of unemployment in the first period after graduating from high school by 12.5 percent and increase the accepted wage on that first job by one-third. Given a greater ability to borrow to finance unemployment spells, agents will hold fewer assets throughout their life cycle; in the experiment, asset holdings would be one-third less 10 years after graduation. Thus, Rendon finds that borrowing constraints importantly affect employment outcomes and asset accumulation.

Paserman (2008): The standard search model is based on the conventional assumption that agents use exponential discounting in weighing the current net cost of search and the future wage payoff from continuing to search. Paserman (2008), following a growing literature in which agents are assumed to have time-inconsistent preferences, specifies and estimates a search model with (quasi) hyperbolic discounting. In addition to allowing for present-biased preferences, Paserman extends the search model to include (i) a decision about search intensity, in essence, a choice about the per-period probability of receiving a job offer, (ii) an exogenous probability of layoff once employed and (iii) the receipt of unemployment benefits for a fixed period of time. The agent solves a finite horizon problem until the point at which unemployment benefits are exhausted and an infinite horizon problem from that point forward. Thus, the reservation wage and search intensity are constant after exhaustion, but are duration dependent during the period when the agent is still eligible for unemployment benefits.

To see the role of hyperbolic discounting, consider the simple discrete time finite horizon search model, abstracting from the additional extensions introduced by Paserman. With hyperbolic discounting, the value functions are:

$$\begin{aligned} V_t &= b + \beta\delta[q\,E\max(W(w), V_{t+1}) + (1-q)V_{t+1}], \\ W(w) &= \frac{w}{1-\delta}, \end{aligned} \tag{175}$$

where as before V_t is the value of searching in period t, $W(w)$ is the value of accepting a wage of w, b is unemployment income net of the cost of search in period t, q is the offer probability, δ is the "long-run" discount factor and β is the "short-run" discount factor. Note that the value of accepting a wage at $t+1$ after searching in period t, as viewed at t, is discounted by δ, that is, exponentially. Thus, it is as if the agent has two selves. The agent who is making a decision in the current period, the current self, is impatient, discounting the expected future payoff to search by $\beta\delta$. However the future self, the self who will receive the benefit of the search and controls future decisions, discounts exponentially. In formulating (175), it was assumed that the current self is aware that when the next period is reached, the current self at that time will be impatient, in which case the agent is deemed sophisticated. This is in contrast to a naive agent, who instead would assume that in the next period the current self would no longer be impatient.

The reservation wage is, as before, the wage that equates the value of search and the value of employment, namely $w_t^* = (1-\delta)V_t$. With a little algebra, we can write (175), analogous to (170), as

$$w_t^* = b(1-\delta) + \beta\delta w_{t+1}^* + \beta\delta h_{t+1}[E(w|w \geq w_{t+1}^*) - w_{t+1}^*]. \tag{176}$$

Obviously, the reservation wage equation with hyperbolic discounting is the same as that with exponential discounting if $\beta = 1$. Further, given its recursive structure, it is clear also that the reservation wage path is lower at all t as the degree of impatience is greater, that is, as β is smaller. Thus, the effect of present-bias in agent preferences is to make job acceptance occur sooner, leading to shorter durations of unemployment and lower accepted wages. The future self, however, would have preferred that the current self be more patient.

As noted, Paserman's model is somewhat more complicated. In particular, he allows for a choice of search intensity, which affects the probability of receiving an offer and which is costly. In this setting, an agent has two instruments to minimize the current cost of search, the choice of search intensity and the choice of the acceptance wage. Della Vigna and Paserman (2005) show that with hyperbolic discounting (in an infinite horizon setting) agents will choose a lower search intensity and a lower reservation wage. Because a lower search intensity leads to a lower offer probability, and thus to longer spells, while a lower reservation wage leads to shorter spells, as in the case of the standard search model, the net effect on expected spell duration is, in general, ambiguous. However, DellaVigna and Paserman show that, as in the standard model, log concavity of the wage offer distribution is sufficient for the expected duration of unemployment to fall.

As in the standard finite horizon model, we can use (176) to consider identification. In the standard model, we noted that b and δ could be separately identified with data on at least three periods that include accepted wages. However, identification is no longer possible with the addition of hyperbolic discounting; one cannot separately identify the structural parameters b, δ and β from $b(1-\delta)$ and $\beta\delta$. But, Paserman also has data on UI benefits. In his model, b is a composite of the level of UI benefits, b^{UI}, and the value of search time, b_0.[129] Thus the first term in (176) is $(b^{UI} + b_0)(1-\delta)$. Clearly, cross-section variation in b^{UI} is sufficient to identify δ, which implies that b_0 and β are also identified.[130]

The model is estimated by maximum likelihood, with an extended version of likelihood function (174) to account for unobserved heterogeneity and layoffs.[131] The

[129] Note that the cost of search in his model is isomorphic to the probability of receiving an offer.

[130] The reservation wage, mean accepted wage and hazard rate are all functions of b^{UI}. They can be estimated as nonparametric functions of b^{UI} Paserman also models search after the exhaustion of benefits. In that case, it is assumed that individuals solve an infinite horizon problem.

[131] The model was estimated both for a normal and log normal wage offer distribution. As found by others (for example,Wolpin (1987), the normal distribution assumption proved problematic. Paserman also allows for measurement error for the reasons previously discussed.

estimation method, like all DCDP models, requires iterating between the solution of the DP problem and calculation of the likelihood function. As is standard in the DCDP literature, Paserman presents evidence on goodness-of-fit to evaluate the performance of the model.

The long-run discount factor, when not estimated to be at the boundary, could not be distinguished from one. The short-run discount factor was estimated to be 0.40 for a low-skilled sample, 0.48 for a medium-skilled sample and 0.89 for a high skilled sample. The p-value for a likelihood ratio test of whether the estimates of β were equal to one were less than 0.01 for the first two samples and 0.08 for the third.

Paserman uses the estimates of the model to assess the impact of policy interventions on unemployment search outcomes and on welfare. Measuring welfare in a hyperbolic discounting model is, however, somewhat problematic as there are, in essence, two agents (selves). Paserman adopts as the welfare measure the exponentially discounted utility of the long-run self under the strategy chosen by the hyperbolic self. In an exponential discounting setting, because there is a single agent making optimal choices subject to constraints, any additional constraints must always reduce welfare. With hyperbolic discounting, this is not necessarily the case. Using this welfare measure, Paserman addresses the question of whether it is possible to design a policy that not only improves welfare, but also reduces unemployment duration and lowers government outlays.

Paserman finds that by imposing a fine, equal to the amount of unemployment benefits, on unemployed agents who do not meet a search effort threshold, it is possible to achieve all of these goals. Indeed, Paserman shows (numerically) that there exists a threshold level of search intensity at which unemployment durations fall and for which the increase in agent welfare and the savings in government outlays is maximized. This experiment implies that program in which the search intensity of unemployed workers is monitored not only may reduce the cost of the UI system (subject to the cost of monitoring), but may also improve the welfare of those who are unemployed.

Ferrall (1997): As we have stressed, DCDP models have been used extensively for policy evaluation. In the present context, for example, most empirical applications of the DCDP approach to job search provide a quantitative assessment of the impact of changes in the UI system, such as altering benefits. Many of those models capture some, but not all, of the features of the UI system, often in a somewhat stylized fashion. It is reasonable to suppose that the closer a model mimics UI program rules, the better the model will be in evaluating policy changes. Ferrall (1997) structurally estimates a DCDP model of job search, which integrates all of the major features of the Canadian UI system.[132]

[132] Ferrall also estimates a model for the US, but, because the UI system varies from state to state and is much less generous than the Canadian system, he does not incorporate UI benefits into the analysis. We focus on the Canadian data to highlight the fact that the DCDP approach allows for a detailed representation of UI policy.

Ferrall studies the transition from school to work. In Canada, as in the US, the period of search for one's first job after leaving school is usually not covered by the UI system. Although that spell of search unemployment is not insured, there is still the potential for the UI system, given its structure, to affect search behavior. To understand why, consider the structure of the UI system in Canada relevant during the time period studied by Ferrall. In that system:

1. An unemployed worker who is eligible for insurance must wait 2 weeks after becoming unemployed before collecting benefits.[133]
2. The benefit level depends on the previous wage through a fixed replacement rate (0.60 at the time) The insurable weekly wage is bounded from below by $106 and from above by $530. Thus benefits are $0 if the wage on the previously held job was less that $106, 0.6 times the wage if the wage is between the bounds and $318 for wages at or above $530.
3. To be eligible for UI benefits, a person must have worked on insurable jobs a certain number of weeks in the 52 weeks prior to becoming unemployed. The number of weeks depends on the regional unemployment rate and the person's previous employment and UI history.
4. The number of weeks of benefits depends on the number of weeks worked on the previous job and on whether the individual is eligible for extended benefits, but is capped at 52.

There are two elements of the UI system that would affect search during an uninsured spell. First, because benefits are paid during insured spells, there is an incentive for individuals in an uninsured spell to become employed to be eligible for benefits during future unemployment spells. Thus the UI system reduces the reservation wage in an uninsured spell; further, the reservation wage will be lower the higher are benefits (Mortensen, 1976). On the other hand, because the level of benefits increases with the wage on the prior job, there will be an incentive for someone in an uninsured spell to wait for a higher wage offer, that is, to have a higher reservation wage. This incentive, however, only applies to individuals whose reservation wage would otherwise be below the maximum insurable wage, although the standard search model would no longer apply. Moreover, the magnitude of these incentive effects depend on the probability that an individual will be laid off from future jobs.

The model estimated by Ferrall, aside from the explicit incorporation of UI rules, differs from the standard single spell search model in a number of ways. The model allows for a search period during school, an initial uninsured spell after leaving school, the first job spell and a subsequent insured unemployment spell if a layoff occurs. The individual

133 Actually, this waiting period is only for those who are unemployed through layoff. Those who quit or were fired had a waiting period of five weeks. Ferrall assumes the waiting period to be two weeks independent of the reason for the unemployment spell.

maximizes the expected present value of the log of consumption, where consumption equals the wage while working and the sum of UI benefits net of the cost of search plus the value of home production. In each period of unemployment, the individual receives a job offer with some positive probability. However, a job offer comes not only with a wage offer, but also with a layoff rate. The wage offer function is Pareto and the layoff rate can take on a fixed number of values, randomly drawn.[134] Individuals differ, according to their unobserved type, in their market skill level and in their value of home production.

The solution method is by backwards recursion, where the value function for the infinite horizon search problem when UI benefits are exhausted after a layoff occurs serves as the terminal value function for the insured unemployment spell at the time of benefit exhaustion. The model is solved backwards from there as a finite horizon problem until the beginning of the search period while in school. The estimation is by maximum likelihood. Ferrall provides evidence of model fit.

Ferrall performs a number of counterfactual experiments that vary the parameters of the UI system. The most extreme is the elimination of the UI system, an out-of-sample extrapolation that is only possible within the structural framework. The resulting impact on unemployment durations depends on geographic location and education.[135] Recall from the earlier discussion that there was no unambiguous prediction of how reservation wages would be affected by such an experiment. Ferrall finds that for those with at most a high school education residing outside of the Atlantic region, reservation wages rise; the expected duration of unemployment after leaving school (including those who have no unemployment spell) is estimated to increase by about 50 percent. Similarly, for those with some college residing outside of the Atlantic region, the increase is 40 percent. However, there is almost no effect on the expected duration for those residing in the Atlantic region regardless of education.

4.3. Dynamic models of schooling and occupational choices

This section describes the use of DCDP models in labor economics to study schooling and occupational choice and to analyze the effects of policy interventions aimed at increasing skill investment, such as tuition and school attendance subsidies and student loan programs. We begin with a brief discussion of the foundational schooling and occupational choice models from the early literature, which tended to be either static models or life-cycle models without uncertainty. These first generation models were influential in shaping the questions addressed and models developed in the later DCDP literature. We then describe contributions to the more modern DCDP literature.

134 As previously noted, the lower bound of the support for the Pareto distribution cannot be identified. Ferrall fixes that value. Christensen and Kiefer (2009) also use the Pareto distribution and impose, based on the wage posting model of Burdett and Mortensen (1998), the individual's reservation wage as the lower bound.

135 UI system parameters depend on region and some of the structural parameters are allowed to differ by education.

4.3.1. Foundational literature

One of the earliest discussions of the determinants of schooling and occupational choices is given by Walsh (1935), which describes a model in which individuals invest in education until the return on education equals the return on other possible investments.[136] The paper also examines the empirical support for the model using data from a variety of sources.[137] Walsh (1935) calculates the returns associated with different levels of schooling and with a subset of occupations (doctor, lawyer, engineer), adjusting for costs (tuition, room and board) and foregone earnings. He raises the potential problem of ability bias in comparing lifetime earnings streams of different education levels and different professions. After finding that the wage returns to being a college graduate and to being a lawyer greatly exceed costs, whereas the returns to receiving an M.A. and Ph.D. are lower than the cost, he attributes the difference to nonpecuniary benefits associated with working in academia.

Roy (1951), in a seminal paper, provides the modern framework for modeling occupational choice as an earnings maximization problem that he then uses to analyze the implications of self-selection into occupations for earnings distributions. The Roy model assumes that individuals are endowed with two different skills, drawn from a joint log normal distribution. Each skill is productive in only one of two occupations, denoted by Roy as hunting and fishing. Skill is measured in units of output produced. Thus, an individual's earnings in an occupation are the product of the price of a unit of occupation-specific output times the amount of skill (output production) embodied in the person; individuals choose to work in the occupation that maximizes their earnings. Roy (1951) did not apply the model to data, but showed that the structural parameters of the underlying model, the means, variances and covariances of the joint skill distribution, can be recovered from earnings data, even though earnings in a particular occupation is only observed for people who chose that occupation. The identification of these structural parameters derives from the theoretical formulation of the determination of earnings and from the distributional assumption.[138]

The literature started by Roy (1951) emphasized the importance of self-selection, skill heterogeneity and latent skills in understanding occupational choices and earnings. However, in Roy's formulation, skills are treated as endowments. Another branch of the literature, associated with Mincer (1958), Becker (1964, 1967) and Ben-Porath (1967), evolved with the aim of understanding the human capital investment (or skill acquisition) decision and the implications for lifetime earnings of acquiring skills through

136 Walsh (1935) is cited by Becker (1964), which is perhaps the most influential work in the development of the human capital literature.

137 The sources include a survey of 15,000 former members of the Alpha Kappa Psi fraternity on the education and earnings, a survey of Land Grant colleges made by the US Department of Interior, published figures on the earnings of physician and doctors, and a survey of Harvard Law graduates.

138 Heckman and Honore (1990) derive additional implications for earnings distributions and provide identification results in the case of non-normal distributions.

schooling and job training investments. Mincer (1958) proposes a lifetime earnings model where the only cost of schooling is foregone earnings. In his model, all individuals are assumed to be *ex ante* identical, which implies a compensating earnings differential for individuals who spend more time in school. In equilibrium, everyone is indifferent between alternative schooling levels, because (discounted) life-time earnings are the same, but there is an earnings premium to each additional year of schooling at every post-schooling age. By equating lifetime earnings for individuals with different levels of schooling, Mincer (1958) derives a log earnings equation that is linear in years of schooling. Mincer (1974) augments the schooling model with a model of on the job investment that leads to a log earnings function that is linear in schooling and quadratic in work experience. That equation has come to be known as the Mincer earnings function, which has had widespread application in empirical work.

Mincer's schooling model is silent about which individuals invest in schooling. Becker (1967), in his Woytinsky lecture, specifies a human capital production function in which the marginal return from investing in human capital declines with investment due to an individual's limited capacity. The marginal cost of investing in human capital, the cost of financing additional human capital investment, depends on access to funding (parental and governmental subsidies and loans to education) and is increasing in the level of investment as cheaper sources of capital are used first. The equilibrium level of human capital investment equates the marginal return to marginal cost (at the intersection of the downward sloping demand curve and the upward sloping supply curve). Becker's model implies that the level of human capital investment will in general differ across people, because of differences they face in either supply or demand conditions. For example, a higher level of innate ability implies a higher demand curve, because higher ability is assumed to make human capital investments more productive.

Rosen (1977) translates this framework into a schooling choice model. The log of earnings is assumed to be a function of schooling and ability. Following Becker, schooling (time spent investing in human capital) increases the stock of human capital and thus earnings, but at a decreasing rate. The marginal return is the derivative of the log earnings function with respect to schooling. The marginal cost of schooling, the interest rate at which an individual can borrow depends on family background. The optimal level of schooling is found by equating the marginal return to the marginal cost.

These previous authors model the human capital investment decision as a one-time decision. Ben-Porath (1967) extends the optimal human capital investment decision to a life cycle setting. The Ben-Porath (1967) model assumes that individuals choose a human capital investment profile to maximize discounted lifetime earnings. Human capital is produced at any age through the application of time (a fraction of an individual's human capital stock) and purchased inputs, conditional on an individual's ability and existing stock of human capital. The fraction of the human capital stock not used to produce additional human capital is used to produce earnings. Similar to Roy (1951),

an individual's earnings at any age is the product of a market determined price of a unit of human capital and the individual's stock of human capital not used in investment at that age. Schooling, in this framework, is viewed as a period of full-time investment (no earnings) and on-the-job training as a time of partial investment. Given a finite lifetime, the optimal human capital investment profile, the fraction of time spent investing, declines with age. Thus, any period of full-time investment, that is schooling, would come first.

Willis and Rosen (1979) empirically implement a model of schooling choice that combines the essential features of this early theoretical literature.[139] The paper develops a two sector model where individuals decide whether or not to attend college, basing their decision on expected lifetime earnings with and without college, on financing capacities that differ by family background and on nonpecuniary benefits of education. The model incorporates two unobservable abilities, associated with high school and college level skills. Willis and Rosen (1979) find that the decision to attend college is strongly influenced by expected lifetime earnings gains and that family background is an important determinant of college-going decisions. In addition, they find comparative advantage to be an important feature of the labor market; that is, high school graduates have better prospects as a high school graduate than would an average college graduate and college graduates having better prospects as a college graduate than would an average high school graduate.

4.3.2. DCDP models

The DCDP literature extends this earlier work on schooling and occupational choice to a dynamic setting, in which individuals face a sequential decision problem with uncertainty. It incorporates features from the earlier literature, allowing for worker heterogeneity, multiple skill types, latent skills, self-selection and comparative advantage. The literature can be broadly categorized into partial equilibrium approaches, which take skill prices as given, and market equilibrium approaches, where there is an explicit link between the prices paid to skill in the economy and aggregate skill quantities. Here, we first describe partial equilibrium models of schooling and occupational choices and then the more limited set of market equilibrium models. Subsequently, we consider DCDP models that have been developed for particular contexts, for example, to analyze the decision about college major or the decision to enter and exit the teaching profession.

Partial equilibrium models of schooling and occupational choice

Gotz and McCall (1984), one of the pioneering papers in the DCDP literature, as noted previously, develops an occupational choice model for the purpose of studying

139 Much of the empirical literature prior to Willis and Rosen (1979) was concerned with estimating rates of return to schooling. A considerable amount of effort was (and still is) devoted to accounting for bias in the schooling coefficient due to omitted ability in a Mincer-style earnings function. See Card (2001) and Wolpin (2003) for an assessment of that literature.

the retention decision of Air Force officers, in particular, how retention responds to compensation policy and to the retirement system. In the model, officers make a binary choice at each age about whether to stay or leave the Air Force so as to maximize the expected present value of pecuniary and nonpecuniary returns. There is a single taste shock that is realized each period and that affects the value of the military option. An officer who leaves joins the civilian labor force and earns a civilian wage. In addition to considering compensation and pension benefits, the model also explicitly accounts for the effects of the chance of promotion on the expected value of staying in the military. The probabilities of promotion and military and civilian earnings are treated as exogenous.[140] The model also allows for persistent differences among individuals in their preference for military service (permanent unobserved heterogeneity).

The model parameters are estimated by maximum likelihood using data on officer employment histories from 1973-1977 as well as data from the Current Population Survey used to construct estimates of civilian earnings. As a way of validating the model, the estimated model is used to forecast retention rates for data not used in the estimation, which shows that the estimated model produces good out-of-sample forecasts. The fit of the dynamic retention model is also compared to that of two competing models, one that does not allow for unobserved permanent preference heterogeneity and a lifetime model without per-period shocks, where individuals know with certainty the year they will leave the military. The dynamic model that allows both for permanent unobserved heterogeneity and for per-period shocks provides the best fit to the data. The dynamic model is then used to assess the effects of a number of policy interventions of interest to the Military, including (i) an increase in pay and allowances, (ii) the introduction of a bonus based on years of service completed, (iii) a decline in the value of the military retirement annuity, (iv) an increase in flight pay and (v) indexing pay to the CPI.[141]

Miller (1984), another pioneering paper in the DCDP literature, develops and estimates a matching model of occupational choice. The model assumes that the pay-off to a particular job within an occupation depends on a match-specific component and a random component. Individuals do not know the match-specific component prior to starting the job, but they observe their output. Beliefs about the quality of the match change with experience on the job through a Bayesian updating procedure. Jobs for which the expected return stream are identical are defined as being the same occupation. The model has implications for which jobs should be sampled first and for how long. For example, the notion of equalizing differences would imply that jobs with high informational content pay less on average in equilibrium and attract relatively inexperienced workers who quickly discover their personal match quality and leave in the

[140] Specifically, pay grades and promotions are assumed to be probabilistic functions of observable state variables.

[141] A related study by Daula and Moffitt (1995) uses a similar DCDP model to analyze the effect of financial incentives on Army infantry reenlistments.

event of a bad match. These types of jobs would include a large number of inexperienced workers in the process of learning about their match plus a small number of experienced, permanent workers. Jobs with lower informational benefits should have less turnover, pay more for new entrants and have a less variable wage distribution.

The dynamics in the model arise due to the learning process as the value of remaining in a job provides not only an immediate payoff but also information about the future payoff in that job. The parameters of Miller's (1984) job-matching model are estimated by maximum likelihood using data on job tenure and job changes, where the discrete time hazard of remaining in a given job or switching to a new job are derived from solving the dynamic programming model.[142] The hazard function is assumed to also depend on the initial observed demographic characteristics of the individual. To capture unobservable heterogeneity, the hazard model depends on two unobserved states, following the approach of Heckman and Singer (1984). The Coleman-Rossi data set, which surveyed a sample of men about their entire work history, education and family background, is used to estimate the model. The empirical evidence supports the prediction that young inexperienced individuals receive low wages in part because they seek out different kinds of occupations, those with greater informational content, than do older individuals.

The preceding papers focused only on the occupational choice decision. The first DCDP model to combine schooling, working and occupational choices in a single framework is Keane and Wolpin (1997). To illustrate concretely the specification of a DCDP model of human capital accumulation, we next describe Keane and Wolpin's (1997) model's structure in detail. As further discussed below, a number of papers in the recent DCDP literature extend the Keane and Wolpin (1997) modeling framework to incorporate additional features.

In the baseline model presented in Keane and Wolpin (1997), individuals make repeated choices over time, starting at age 16 and ending at age 65, about whether to participate in one of five different sectors of the economy: (i) work in a white-collar occupation, (ii) work in a blue-collar occupation, (iii) work in the military, (iv) attend school, or (v) engage in home production. There is a finite horizon during which individuals accumulate schooling and occupation-specific experience that affect future wage earning opportunities.

Denote the five choice alternatives in each time period by m where $m \in \{1, 2, 3, 4, 5\}$. The first three alternatives $\{1, 2, 3\}$ are the work alternatives, the 4th is the schooling alternative and the last is the home alternative. Let $d_{ma} = 1$ if alternative m is chosen in time period a. R_{ma} represents the reward (contemporaneous utility) from choosing alternative m, which captures all benefits and costs associated with that alternative.

[142] The model is solved using the Gittens index, a simplification in the solution of the dynamic programming problem that arises because tenure in one job does not affect the rate of learning in other jobs.

The reward in a work sector is the wage, which is the product of the price paid per unit skill times the amount of skill accumulated in that occupational sector. Let r_m denote the rental price paid to skill in occupational sector m and e_{ma} the occupation-specific skill units.

$$R_{ma} = w_{ma} = r_m e_{ma}, \quad m \in \{1, 2, 3\}. \tag{177}$$

The technology for skill production depends on the number of years of schooling, g_a, and on occupation-specific work experience, x_{ma}. The production function takes the form:

$$e_{ma} = \exp[e_{m,16} + e_{m1} g_a + e_{m2} x_{ma} - e_{m3} x_{ma}^2 + \varepsilon_{ma}], \tag{178}$$

where $e_{m,16}$ represents the endowment of skill at age 16. The log wage equation is:

$$\ln w_{ma} = \ln r_m + e_{m,16} + e_{m1} g_a + e_{m2} x_{ma} - e_{m3} x_{ma}^2 + \varepsilon_{ma}. \tag{179}$$

The wage equation has the Mincer form of being linear in years of education and quadratic in experience but has the Ben-Porath (1967) and Griliches (1977) pricing equation interpretation.

If a person goes to school, the per period reward is:

$$R_{4a} = e_{4,16} + \varepsilon_{4a} - t_{c1} 1(\text{attendcollege}) - t_{c2} 1(\text{attendgraduateschool}), \tag{180}$$

where t_{c1} and t_{c2} are tuition costs, $e_{4,16}$, is endowed skill at age 16 and ε_{4a} is a random shock component. The home alternative has the associated nonpecuniary reward:

$$R_{5a} = e_{5,16} + \varepsilon_{5a}, \tag{181}$$

where $e_{5,16}$ is the skill endowment and ε_{5a} the random shock component.

The initial conditions in the model are the highest grade completed at age 16 ($g(16)$) along with the unobserved skill endowments in the different sectors. It is assumed that accumulated experience is zero for all alternatives in the first period. The shock components are assumed to be joint normally distributed and serially independent, conditional on the unobserved endowments:

$$\begin{pmatrix} \varepsilon_{1a} \\ \varepsilon_{2a} \\ \varepsilon_{3a} \\ \varepsilon_{4a} \\ \varepsilon_{5a} \end{pmatrix} \sim N(0, \Omega). \tag{182}$$

The state vector at any a is described by

$$\Omega_a = \{e_{16}, g_a, x_a, \varepsilon_a\}, \tag{183}$$

where

$$e_{16} = [e_{1,16}, e_{2,16}, e_{3,16}, e_{4,16}, e_{5,16}], \tag{184}$$

is the vector of age-16 endowments,

$$x_a = [x_{1a}, x_{2a}, x_{3a}, g_a] \tag{185}$$

is vector of work experience and schooling accumulated in the different sectors and

$$\varepsilon_a = [\varepsilon_{1a}, \varepsilon_{2a}, \varepsilon_{3a}, \varepsilon_{4a}, \varepsilon_{5a}] \tag{186}$$

is the vector of shocks.

The value function at age a is the maximized value of the expected remaining lifetime utility, taken over all possible sequences of future choices, with respect to the choice at a,

$$V_a(\Omega_a) = \max_{\{d_{ma}\}} E\left[\sum_{t=a}^{A} \delta^{\tau-a} \sum_{m=1}^{5} R_{ma} d_{ma} | \Omega_a\right]. \tag{187}$$

The problem can be written recursively in Bellman equation form. For $a < A$, the alternative specific value function is

$$V_{ma}(\Omega_a) = R_{ma}(\Omega_a) + \delta E\left[V_{a+1}(\Omega_{a+1}) | \Omega_a, d_{ma} = 1\right], \tag{188}$$

where the expectation is taken over the random shock components. In the last time period, A,

$$V_{mA}(\Omega_A) = R_{mA}(\Omega_A). \tag{189}$$

The value function is the maximum over the alternative specific value functions:

$$V_a(\Omega_a) = \max_{m \in M} V_{ma}(\Omega_a). \tag{190}$$

The state variables that evolve over time are the accumulated sector-specific experience and the years of completed schooling:

$$x_{m,a+1} = x_{ma} + d_{ma} \quad m = 1, 2, 3 \tag{191}$$

$$g_{a+1} = g_a + d_{4a} \quad g(a) \leq \bar{g}. \tag{192}$$

The observed data are the sector choices that people make and their observed wages (for the sectors with pecuniary rewards) starting from age 16 and ending at age $\bar{a}$ (at most age 27 in the data):

$$[d_{ma}, w_{ma}d_{ma} : m \in \{1, 2, 3\}] \quad [d_{ma} : m \in \{4, 5\}]. \tag{193}$$

It is assumed that individuals observe contemporaneous shocks ε_a, but that the researcher does not. The observed state space (exclusive of the shocks) is

$$\Omega_a^- = [e_{16}, g_a, x_a]. \tag{194}$$

The likelihood is

$$\Pr[c_{16}, \ldots, c_a | g_{16}, e_{16}] = \prod_{a=16}^{\bar{a}} \Pr[c_a | \Omega_a^-], \tag{195}$$

where c_a denotes the vector of choices and wages at age a. The estimation proceeds by: (i) choosing an initial set of parameters, (ii) solving the dynamic programming problem numerically (by approximating the $E\max_a$ functions as previously described), (iii) computing the likelihood, and (iv) iterating to maximize the likelihood until convergence.

The baseline model that Keane and Wolpin (1997) estimate also includes unobservable heterogeneity. Specifically, there are assumed to be 4 types of individuals with heterogeneous age 16 endowments, denoted by

$$e_{k,16} = \{e_{mk,16} : m = 1, \ldots, 5, k = 1, \ldots, 4\}. \tag{196}$$

The type of the individual is assumed to be known to individuals but unknown to the researcher. Unobservable heterogeneity introduces the potential for comparative advantage into the model in that some individuals persistently get higher rewards in certain sectors, but perhaps not others. Unobserved permanent endowment differences are necessary to fit the high degree of persistence in choices observed in the data.

In the model, the only observable initial conditions that varies is schooling attained at age 16, g_{16}.[143] If the shocks were serially correlated, then it would be problematic to condition the analysis on g_{16}, because g_{16} likely reflects prior schooling decisions that would be affected by earlier shocks. If the shocks are iid, however, then conditioning the analysis on g_{16} is not problematic. The maintained assumption is that the initial condition g_{16} is exogenous with respect to the shocks conditional on the unobserved type.

[143] Work experience, $x_{16} = 0$.

Accounting for unobservable heterogeneity and for initial conditions, the likelihood is:

$$\prod_{n=1}^{N} \Sigma_{k=1}^{L} \pi_{k|g_{n,16}} \Pr[c_{na}|g_{n,16}, type = k]. \tag{197}$$

The type probability is estimated as a function of the initial schooling.

Estimation of the model's parameters is based on eleven years of data on young white males from the NLSY79. The analyses subsample consists of 1373 observations on white males who were age 16 or less as of Oct. 1, 1977 and who are followed through 1988. Each time period in the model corresponds to one year in the data. Wages are measured as full-time equivalent wages, estimated as average weekly wages times 50. Parameter estimates are obtained by simulated maximum likelihood, as previously described.

Keane and Wolpin (1997) evaluate the goodness-of-fit of their baseline model (described previously) and of a few alternative model specifications that differ in their degree of parsimony to learn which features of the model are important to achieving a good fit. Their preferred model augments the baseline model to incorporate skill depreciation during periods of nonwork, job-finding costs, school reentry costs, and nonpecuniary components of work sector alternatives. As a way of validating the model, Keane and Wolpin (1997) also evaluate the fit of the model out-of-sample by predicting the choices of younger birth cohorts (using CPS data) that were not used in estimating the model.

A consistent empirical finding (Willis and Rosen, 1979; Heckman and Sedlacek, 1985) is that comparative advantage plays an important allocative role in the labor market. Workers self-select into occupations and into sectors based on their relative productivities. Keane and Wolpin (1997) find that comparative advantages determined by age 16 lead to large differences in school attainment and later labor market outcomes. Indeed, most of the variation in lifetime utility comes from inequality in skill and preference endowments at age 16, pointing to the importance of early influences in explaining lifetime inequality.

The estimated model is used to predict the effects of a $2000 (1987 dollars) college tuition subsidy on the college going rate. Under the preferred model specification, the subsidy increases the high school graduation rate by 3.5 percentage points and the college graduation rate by 8.5 percentage points. However, the main beneficiaries of the subsidy, in terms of lifetime utility, are individuals who would have gone to college without the subsidy.

A follow-up paper by Keane and Wolpin (2000) uses a similar framework to analyze the sources of black/white differentials in schooling attainment and earnings and to assess the impact of policies intended to close the racial gaps. Race enters the model in a number of ways, as a determinant of preference parameters, unobserved type probabilities, and wages. The paper finds that differences in initial age 16 skill endowments are the primary explanation for low schooling attainments of blacks relative

to whites. This finding has important implications for policy. Keane and Wolpin (2000) implement a scheme to equalize the schooling distributions of black and white males through the combined use of a high school graduation bonus and a college graduation bonus. Although this policy, by design, closes the racial schooling gap, it has only a very small effect on the racial earnings gap due to differences in skill endowments at age 16.

An area of research that has received much attention in the nonstructural literature focuses on the effect of credit market constraints on college enrollment. The finding in that literature that tuition effects are inversely related to parental income has often been interpreted as evidence for the existence of borrowing constraints that have adverse consequences for college attendance (see, e.g., Kane, 1999, p. 63). A paper by Keane and Wolpin (2001) studies how borrowing constraints and parental transfers affect educational attainment by estimating a DCDP model of schooling, work and savings decisions of young men, using data from the NLSY79 cohort. The model allows for parents to provide transfers to youths, which the youths take as given and which vary depending on whether the youth chooses to go to college. Like the previous papers, the model incorporates unobserved heterogeneity (endowments at age 16). In the model, schooling and work are not mutually exclusive choices and youths can work full or part time while still attending school (full or part time). Youths may borrow up to a limit. Keane and Wolpin (2001) find that borrowing constraints are tight (financing college tuition through uncollateralized borrowing is not feasible in the model). In addition, consistent with the pattern found in the nonstructural literature, Keane and Wolpin report that a tuition increase generates a pattern of larger percentage declines in enrollment for youth whose parents have lower SES.

On the surface, it would appear that the inference drawn in the nonstructural literature, that borrowing constraints exist and limit college attendance of youths from less affluent families, is validated by the congruence of these two findings. However, when Keane and Wolpin simulate the impact of relaxing the borrowing constraint, by allowing youths to borrow the full tuition cost, they find that there is only a negligible increase in college attendance. However, allowing college attendees to borrow up to the full tuition amount leads to a reduction in their propensity to work while attending school and to an increase in their consumption. They therefore conclude that college attendance is not limited to any great extent by borrowing constraints, but rather primarily by age-16 endowments of pre-market skills and/or preferences.[144]

[144] Keane and Wolpin (2001) also find that, on average, youths receive a transfer from their parents in excess of what is received when not attending college, sufficient to fully subsidize college tuition costs. The subsidy ranges from about one-half of the tuition cost for youths whose parents are the least educated (neither a high school graduate) to almost twice the tuition cost for youths whose parents are the most educated (at least one parent a college graduate). It might appear that it is because of the largesse of parents that relaxing borrowing constraints has only a minimal impact on college attendance. However, simulating the impact of relaxing the borrowing constraint in a regime where parents are assumed to provide no additional transfers to children who attend college leads to the same result. Transfers do, however, have a non-negligible effect on school attendance. If transfers are equalized across children, high SES children go to school less, but low SES children do not increase their attendance by much.

The finding that borrowing constraints are tight and yet relaxing them does not lead to increased college attendance has been controversial. However, it is consistent with earlier research by Cameron and Heckman (1998, 1999) that estimates a sequential model of school attendance decisions.[145] That research finds a strong positive correlation between family income and college attendance, conditional on high school graduation, even after controlling for effects of dynamic selection on unobservables. After controlling for AFQT test score (interpreted as a proxy for the individual's endowment at age 16), however, liquidity constraints no longer play a strong role in college attendance decisions.

In the previously described papers, log wages are specified as a linear function of the number of years of schooling. Belzil and Hansen (2002) estimate a DCDP model of schooling decisions with a focus on allowing the returns to different levels of schooling to vary. In particular, they model the wage equation as a spline in years of schooling with eight knots. Their model assumes that individuals make sequential decisions as to whether to attend school for up to 22 years, after which they enter the labor market.[146] While in school, they receive parental transfers according to a parental transfer function that depends on accumulated years of schooling. After entering the labor market, individuals are employed with some probability and, if employed, receive a wage rate. Both the wage rate and the probability of employment depend on their schooling attainment and labor force experience.[147] The model has three source of uncertainty: a schooling preference shock, a wage shock and an employment shock. It also includes six unobserved types to capture unobservable heterogeneity in schooling ability and in market ability, and one of the goals of the paper is to recover the correlation between unobserved schooling and market ability.

The model is estimated by simulated maximum likelihood on a sample of white males from the NSLY79. The estimated parameters indicate that log wages are convex in years of schooling, with statistical tests rejecting the hypothesis of linearity. The log wage equation has estimated returns to schooling that are very low (1%) until 11th grade, increase to 3.7% in grade 12, and exceed 10% between grades 14 and 16. The estimated returns to schooling are substantially lower than corresponding OLS estimates. For a linear in schooling specification, one obtains an OLS estimate of 10%, in comparison with the structural model estimates of 2% on average up to grade 12 and 7% after grade 12.[148] They also find that there is a strong positive correlation (0.28) between market

145 The approach taken by Cameron and Heckman (1999) can be interpreted as estimating the approximate decision rules from a DCDP model.

146 All individuals are assumed to complete at least six years.

147 In the model, there is also a probability of experiencing a so-called interruption, which is a decision period when no decision is made and the stock of accumulated human capital remains fixed, intended to capture an event such as illness or academic failure that lasts one period.

148 See Belzil and Hansen (2002) Table VIII. The previously described Keane and Wolpin (1997) model did not allow for nonlinearities in returns to education but did allow the return to differ between white and blue collar occupations and found a much higher return in the white collar occupation, which could be viewed as consistent with Belzil and Hansen's (2002) finding of high returns at higher education levels.

ability and realized schooling, which would imply that estimated returns to schooling from wage regressions that do not control for the endogeneity of schooling will tend to be upwardly biased.

Sullivan (2010) develops a DCDP model that combines a model of labor force dynamics (as in Wolpin (1992) and Rendon (2006)) with a human capital model of schooling and occupational choice (as in Keane and Wolpin (1997)). The previous literature considered job search as a separate phenomenon from schooling and occupational choice, though the choices are clearly related. In Sullivan's (2010) model, workers decide in each period whether to attend school and/or work in one of five occupations or neither work nor attend school. An individual who has not graduated from high school may also decide to earn a GED. An employed individual may stay at the current job or switch jobs either within the same occupation or with a change in occupation. Human capital accumulated through work experience is both firm- and occupation-specific. Individuals have heterogeneous skill endowments and preferences for employment in different occupations. Wage offers include a match-specific component, reflecting worker-firm permanent match productivity, and an iid time varying shock. Search arises because of variation in worker-firm match productivity together with mobility costs. Model parameters are estimated by simulated maximum likelihood using data from the NLSY79.

The model estimates are used to perform a number of counterfactuals. Sullivan's (2010) analysis finds that occupational and job mobility are critical determinants of life cycle wage growth, quantitatively more important than the accumulation of occupation-specific human capital. As in previous research, the results also indicate the importance of comparative advantage in understanding schooling and occupational choices. Sullivan also finds that unobservable heterogeneity plays a relatively smaller, though still substantial, role in explaining labor market outcomes than has been found, for example, by Keane and Wolpin (1997). He attributes 56% of the variation in lifetime utility to permanent heterogeneity, which compares to 90% found in Keane and Wolpin (1997).

General equilibrium models of schooling and occupational choice

Most of the literature on modeling occupational and schooling choices is partial equilibrium, taking skill prices as given. However, a few papers in the literature estimate general equilibrium models in which skill prices respond to changes in aggregate market demand and supply for skills.

The earliest paper to estimate a multi-sector general equilibrium model is Heckman and Sedlacek (1985). The paper is an extension of Roy (1951). Although static, and thus not a DCDP model, the paper serves as a link to the later general equilibrium models that fall within the DCDP paradigm. Specifically, Heckman and Sedlacek (1985) estimate a model of individuals' decisions among three sectors: work in the manufacturing sector, work in the non-manufacturing sector or not work. In addition to specifying the

micro-level supply-side sector choice model, the paper estimates an aggregate demand function for skill. The micro supply-side model and the aggregate demand models are used jointly to simulate the effects of price changes on employment levels and wages, such as an increase in the price of energy that predominantly affects labor demand in manufacturing.

Establishing the link between aggregate skill quantities and skill prices can be conceptually important for analyzing the effects of policy interventions. To illustrate, consider, for example, the impact of a college tuition subsidy on the fraction of people going to college. A tuition subsidy must act as a positive inducement to college attendance. In a general equilibrium framework, a college tuition subsidy that induces more people to go to college would also decrease the price paid to college skill given the increase in the aggregate quantity of college educated labor. For this reason, we would expect the predicted general equilibrium effect of a college tuition subsidy on college-going to be smaller than the predicted partial equilibrium effect. The quantitative significance of the supply effect on skill prices is an empirical question.

The papers described below develop and estimate general equilibrium models incorporating schooling and employment choices. The goals of these papers are to understand historical wage and employment patterns for workers of different skill levels and to analyze the effects of skill formation policies, such as tuition subsidies.

Wage inequality has increased at least since the 1980's, with low skill workers experiencing both absolute and relative declines in real wages as the economic returns to skill acquisition have risen. Heckman et al. (1998) (HLT) present the first general equilibrium model of schooling and job training choices, which they use to explore alternative explanations for observed wage patterns and to simulate the effects of college tuition subsidies. In the HLT model, individuals make decisions about whether to go to college, about post-school on-the-job training (human capital investments a la Ben-Porath) and about life-cycle savings. The model assumes that individuals decide whether or not to go to college and on their optimal life cycle consumption and human capital investment paths, assuming they work each period until the age of retirement.[149] There are no credit constraints. The market wage for each skill type is the product of the skill rental price and the amount of accumulated skill. Individuals are heterogeneous in terms of initial skill endowments, captured by the observed AFQT test score.

The model is solved for overlapping generations of agents and estimated using both aggregate CPS data (from 1963 to 1993) and longitudinal data from the NLSY79. Cohorts make different choices because they face different (known) skill prices over their lifetimes. The model assumes a one-to-one correspondence between schooling groups (high school and college) and skill types, that is, that different skill types cum schooling classes are imperfectly substitutable. However, age groups within a given schooling group

[149] The model is life cycle rather than dynamic in the sense that new information is not revealed to the agent in each decision period.

(high school or college) are perfect substitutes. Skill prices are determined in equilibrium. Equilibrium skill prices induce aggregate skill supplies that equate marginal revenue skill products to skill prices.

HLT calculate the partial equilibrium and general equilibrium impact of a 100 dollar increase in tuition on college enrollment. They find that the partial equilibrium response is a decline in enrollment of 1.6 percent. However, when they allow for skill prices to adjust to the reduction in college skill, that is for the increase in the relative price of college skill, they find that the decline in enrollment is less than 0.2 percent. Thus, the adjustment in the relative price of college to high school skill almost completely offsets the disincentive to acquire schooling. Presumably, a tuition subsidy of a similar magnitude can be expected to lead to only a negligible increase in college enrollment due to the fall in the relative college skill price.

Lee (2005) estimates an alternative formulation of a general equilibrium schooling and occupational choice model. The specification of the individual's problem parallels that of Keane and Wolpin (1997). Specifically, in each period an individual decides whether to attend school, work in one of two occupations, blue collar or white collar jobs, or do neither. Individuals are heterogeneous in skill and preference endowments and are subject to idiosyncratic time-varying shocks. A critical difference between the model of Lee (2005) and that of HLT (1998) is that in Lee's model, an individual's skill type is not equated with their schooling. Schooling augments both white and blue-collar skill, though differentially, and it is the aggregate levels of the occupation-specific skills that enter as inputs into the aggregate production function. In Lee's (2005) model, occupations are not perfectly substitutable, but education types are perfectly substitutable within occupation and age groups. This difference between the HLT (1998) and Lee (2005) models has important consequences for the relationship between partial and general equilibrium effects of tuition policies.

Lee (2005) estimates the model using simulated method of moments applied to CPS data on schooling, occupational choice, employment and cohort size, under an assumption that individuals have perfect foresight about future skill prices. The estimated model is then used to investigate how cohort size affects skill prices and wages and also to evaluate the effects of a college tuition subsidy. As in HLT, Lee evaluates the partial and general equilibrium impacts of a 100 dollar increase in tuition on college enrollment. The partial equilibrium effect ranges from 1.2 to 1.9 depending on age and gender, similar in magnitude to HLT. However, the general equilibrium effect is only found to be about 10 percent lower, in sharp contrast to the result in HLT. The reason, supported by simulations performed by Lee, is that workers can respond to changes in the relative price of blue- and white-collar skill by switching sectors. A tuition increase that reduces college enrollment, and thus increases the relative white collar skill price, induces some blue-collar workers to switch to white-collar jobs, mitigating the rise in the white collar skill price.

A recent paper by Lee and Wolpin (2010) develops and estimates a general equilibrium model to explain the evolution of wages and employment over the last 30 years, including gender differentials in employment and earnings, which were not considered in the previous two studies that focused only on males.[150] Specifically, the study aims to account for changes in wage inequality (both overall and within demographic groups), increases in relative wages and employment of women, and a shift that has occurred over time in employment from the goods to the service producing sector. There is an extremely large, mostly nonstructural, literature that considers each of these major labor market changes as separate phenomena.[151] Lee and Wolpin (2010) develop a comprehensive framework which includes many of the factors considered to be potential explanations for these major labor market changes.[152]

The model estimated in the paper has two production sectors, corresponding to goods and services. Aggregate production depends on three skill types (white-, pink- and blue-collar) and on capital. There are time-varying neutral and non-neutral technological changes as well as combined aggregate productivity and relative product price shocks. The goods-to-service product price and the price of capital evolve exogenously.

In the model, men and women age 16-65 can choose to work in any of six sector-occupations (pink collar, white collar or blue collar in either the goods or service sectors), to attend school or to stay home. Each period, individuals receive wage offers from each sector-occupation that depend on schooling attainment and accumulated experience in each sector-occupation. There are also nonpecuniary payoffs and preference shocks to the different options. To capture lower labor force participation rates of women during child-bearing ages, the value of the home choice is assumed to depend on the number of preschool age children in the household. It is also allowed to vary over time to reflect technological improvements that are thought to have occurred in the home sector. In addition, there is a cost of transiting between sector-occupations, which can be interpreted as labor market frictions.[153] The population at any point in time consists of overlapping generations of both sexes. Unobservable heterogeneity is incorporated by including five unobserved types of individuals who differ in sector-specific endowments and in preferences for the home and school options.

150 An earlier paper by Lee and Wolpin (2006) develops a similar model with a focus on examining the relative importance of labor demand and supply factors in explaining the expansion of service sector employment

151 The general rise in inequality and the college premium have often been linked (for example, Murphy and Welch (1992)), but not together with the rise in female-male wages and the growth of the service sector. The growth in the service sector has also been linked with the rise in female employment Fuchs, 1980; Welch, 2000 draws a link between the rise in wage inequality among men and the reduction in the gender wage gap.

152 For a review of the larger literature, see Katz and Autor (1999). The papers they survey include Bound and Johnson (1992), Gottschalf and Moffitt (1994), Juhn et al. (1993), Katz and Murphy (1992), Krusell et al. (2000) and Murphy and Welch (1992, 1993). Recent contributions to this literature include Baldwin and Cain (2000), Eckstein and Nagypal (2004), Hornstein et al. (2005) and Welch (2000).

153 There is a large labor economics literature on interindustry wage differentials among otherwise observably identical workers. Frictions to switching sectors are sometimes proposed to explain these differentials.

Skill prices are equated to marginal revenue products evaluated at aggregate skill amounts. The paper also develops a belief consistent forecast rule for future skill prices, as an approximation of a rational expectations equilibrium. Model parameter estimates are obtained by simulated method of moments, matching the model's predicted levels of wages, employment and school enrollment to data from the CPS, BLS and NLSY79.

Lee and Wolpin (2010) use the estimated model to assess the relative contribution of changing technology, preferences and exogenous forcing variables (the goods to service product price, the price of capital, fertility) as explanations of the previously described major labor market changes. This is done by using the model to simulate labor force outcomes under hypothetical scenarios relative to a baseline economy. The key findings from the analysis are that (i) neutral technological change best accounts for service sector employment growth, (ii) skill-biased technological change best explains the rise in the college wage premium and the increase in overall wage inequality, (iii) the combination of neutral and biased technological change account for the declining gender gap and increased female labor force participation, and (iv) changes over time in fertility and in the valuation of the home sector can account for female-male relative wage and employment growth. The study concludes that a competitive general equilibrium model of the labor market provides a comprehensive framework for analyzing the determinants of wage and employment changes over the last 30 years and that both demand and supply side factors are required to account for the major labor market changes.

4.3.3. The use of DCDP models in related contexts

The previously discussed studies focused on schooling and occupational choice decisions. We next describe a DCDP literature that develops models to study how marriage decisions interact with labor force decisions, the operation of particular occupational labor markets, various behaviors of adolescent youth and the effects of job training on training program participants.

Marriage and career decisions

Gould (2008) estimates a DCDP model of marriage and career decisions of young men age 16-35 using data from the NLSY79 with the aim of exploring the extent to which schooling and employment choices are influenced by marriage market considerations. Individuals choose among four sector options: schooling, white-collar work, blue-collar work and home. In addition, men face potential marriage opportunities that are conditional on their demographic characteristics and on their current and previous marriage, schooling and employment decisions. Based on the available opportunities, they decide among marriage states. Specifically, there is some probability of receiving a marriage offer (from a woman of given type) and men decide whether to accept the offer. Married men also face an exogenous probability of having their marriage terminated by their wife. The model incorporates four unobserved types of individuals to allow for unobservable heterogeneity.

Model parameters are estimated by simulated maximum likelihood.[154] The estimated model is used to study how young men's career choices would change if there were no marriage market returns to career decisions, that is, by shutting down marriage within the model. Simulation results show that the marriage market significantly affects men's schooling and labor market decisions. Without marriage, men work less, study less, and relatively more often choose blue-collar over white-collar work. Another simulation examines the effects of changing divorce costs on men's choices. A decrease in divorce costs leads men to take fewer measures to guard against a marital break-up; they invest less in education and relatively more often choose blue-collar over white-collar work. Overall, Gould (2008) finds that the private returns to human capital investment include significant returns in the marriage market.

Occupational labor markets

We next describe three studies that use DCDP models to analyze the operation of a particular occupational labor market. Sauer (1998, 2004) studies life-cycle career choices of law school graduates following graduation from the University of Michigan Law School and how these choices are affected by financing options and loan forgiveness programs. Stinebrickner (2001) studies the decisions by certified elementary and secondary teachers to stay in or exit from the teaching sector.

Sauer (1998) estimates a model of a law school graduate's choices among five employment sectors that differ in pecuniary and nonpecuniary returns, in promotion and dismissal probabilities, and in the extent to which human capital is transferable across sectors. The possible employment sectors are nonprofit, elite private law firm, non-elite private law firm, separate business and sole proprietor. Attorneys choose among the different sectors taking into account effects of current choices on future job opportunities and wage offers, which depend on endogenously accumulated sector-specific work experience. Within private law firms, lawyers also have the opportunity for promotion from an associate to a partner position. The model includes unobserved types that are assumed to be known to both the worker and to the firm, but not to the researcher. The model is estimated by simulated maximum likelihood.

An interesting aspect of the model is that it generates sector-specific non-monotonic hazards in duration of employment, as observed in the data, through a mechanism that is different from that of the classical job-matching model (Jovanovic, 1979). In Sauer's (1998) model, the ability of the worker and the quality of the match are known from the beginning, and non-monotonic hazards arise because of self-selection. In particular, high ability lawyers face higher probabilities of promotion at private law firms and stay at these firms when they get promoted. Low ability lawyers initially also work at private firms even though they have a low probability of getting promoted, because their experience

154 The maximum likelihood approach, developed by Keane and Wolpin (2001) and extended in Keane and Sauer (2009) allows for classification error and for missing state variables.

pays off later in the form of higher-paying jobs in other sectors. The self-selection mechanism has implications for effects of policy interventions in the market for lawyers, such as programs that forgive loans if a lawyer enters the nonprofit sector. Simulations using the estimated model indicate that a loan forgiveness program induces low ability types to enter the nonprofit sector earlier but is relatively ineffective in attracting high ability lawyers.

A follow-up paper by Sauer (2004) extends his previous DCDP model to incorporate educational financing decisions. The study's goal is to measure the effects of short-term parental cash transfers and family background on educational borrowing and in-school work decisions, and ultimately on earnings after graduation, and also to better understand effects of policies such as tuition tax credits and loan forgiveness programs on these decisions and on post-graduation outcomes. The model assumes that individuals maximize their expected present value of lifetime utility by making decisions on the level of educational indebtedness, whether to work while in school and the type of post-graduation employment. Total financial resources during law school come from five possible sources: parental cash transfers, initial assets, stochastic unobserved assets, educational debt, and stochastic labor income. Post-graduation job market choices are modeled analogously to Sauer (1998). The model also includes three unobserved types to capture unobservable heterogeneity, where the type probabilities depend on family background variables (that include whether the father was an attorney) and whether the individual has an Ivy League BA.

Model parameters are estimated by simulated maximum likelihood allowing for classification error. The estimated model is used to examine the effects on student's borrowing, work while in school and subsequent employment choices of a loan forgiveness program that grants an annual subsidy equal to an individuals' debt obligation for individuals who take jobs in the nonprofit sector within the first 10 years after graduation. Simulations of the model with and without such a program indicate that the program increases borrowing and reduces work while in school. However, the loan forgiveness program has essentially no effect on the choice of first job, with the same types of individuals being most likely to enter the nonprofit sector.[155] The main difference is that they enter that sector with more debt. The effect of the loan forgiveness program on participation in the nonprofit sector is similar to that found in Sauer (1998), except that now allowing for individuals to change their borrowing behavior increases substantially the cost of providing the program.

Stinebrickner (2001) develops and estimates a DCDP model to study the decision of certified elementary and secondary teachers to remain in the teaching sector, to exit into the nonteaching sector or to leave the labor force. Certified teachers often leave the teaching sector within two to nine years following certification. This high turnover is of

[155] These results are similar to those Keane and Wolpin (2001) obtain from simulations of the effect of reducing borrowing constraints on enrollment.

particular concern because certification requirements and wage rigidities in the teacher labor market make it difficult for the market to adapt to fluctuations in teacher demand.

In Stinebrickner's (2001) model, certified teachers receive wage offers in each period in both the teaching and nonteaching sectors. They decide whether to work in the teaching sector, in the nonteaching sector or to not work. The model incorporates marital status and number of children, which are assumed to evolve exogenously. Also, wage offers in both the teaching and nonteaching sectors are allowed to depend on the individual's SAT score, interpreted as a measure of academic ability. Model parameters are estimated by simulated maximum likelihood using data on 450 certified teachers from the National Longitudinal Survey of the High School Class of 1972.

A key result from the analysis is that the primary cause of leaving decisions by teachers is not the relative attractiveness of nonteaching occupations but rather the decision not to work, which for women is strongly influenced by changes in marital status and in numbers of children. For teachers with a high SAT score, though, relatively better options in the nonteaching sector, which has a larger earnings premium for skills measured by the SAT, is a factor influencing their decision to leave the teaching sector. Model simulations indicate that teacher labor supply is responsive to changes in teaching wage offers. Increasing the teacher wage by 20 percent increases the proportion of person-years spent in teaching from 0.5 to 0.8, with a greater response among teachers with higher SAT scores.

Schooling-related choices

There have been a few applications of DCDP models to analyze youth behaviors while in school, for example, the decisions by youth to work while in school, drop-out of school, to enroll in college or to major in a particular subject in college. Eckstein and Wolpin (1999) use a DCDP model to study the determinants of school dropout decisions and to analyze whether working while in school is detrimental to school performance. In the model, youths choose among various work-school combination alternatives so as to maximize expected lifetime utility. Youths who attend high school accumulate credits towards graduation and receive grades reflecting their performance. In each period, they also receive random wage offers for either part-time or full-time employment, which they can either accept or reject. The wage offers depend on their skill endowments, educational attainment and previous labor market experience. Working potentially reduces school performance, as measured by course grades, and thus may increase the probability of failing to progress. The model also incorporates unmeasured heterogeneity at the time of entering high school, in preferences, abilities and in the expected value assigned to receiving a high school diploma. The model is estimated by simulated maximum likelihood using white males from the NLSY79.

Determining the impact of work on high school performance has been the subject of a substantial economics, sociology, and psychology literature.[156]

[156] See Greenberger and Steinberg (1986).

Eckstein and Wolpin (1999) finds that working while in school reduces academic performance, but the quantitative effect is small. A hypothetical policy that forces youths to stay in high school for five years without working or until they graduate increases the percentage of high school graduates by only 2 percentage points, but increases the average number of years of high school completed by dropouts by one year. As in other studies, initial traits at the time of starting high school are found to be major determinants of dropping out behavior. Youths of the types with lower school ability and/or motivation, a lower expected value of a high school diploma, a higher value placed on leisure time, higher skills in jobs that do not require a high school diploma, and a lower consumption value of attending school tend to drop out of school. The implication is that youth labor policies that do not alter the traits that youths bring to high school will be relatively ineffective in improving school outcomes.

Arcidiacono (2005) uses a DCDP model to study how changing the admission and financial aid rules at colleges affect future earnings of individuals. Specifically, he develops and estimates a behavioral model of decisions about where to submit college applications, which school to attend and what field to study. In the model, individuals make application decisions based on their expectation of the probability of acceptance, the application cost, the expected financial aid conditional on acceptance, and on an expectation of how well they will like a particular college and major combination. Schools make admissions and financial aid decisions; but rather than specifying and structurally estimating the school optimization problem, it is assumed that the school's maximization problem leads to a logit probability of a particular student being admitted to school conditional on the quality of the school and the individual's own ability. School quality is measured by the average math and verbal SAT at the school.[157]

Conditional on the offered financial aid and acceptance set, individuals decide what school to attend and what field to study. They also have the option of not attending school and going directly to the labor market. After college, individuals enter the labor market and their expected utility is equated to the log of the expected present value of lifetime earnings. The model is estimated using panel data on high school graduates from a single cohort (the National Longitudinal Study of the Class of 1972). Parameters are estimated by simulated maximum likelihood.[158]

The estimated model is used to examine (i) the effects of affirmative action on college-going decisions of African American students and on their labor market outcomes and (ii) the reasons for large earnings and ability differences across college majors. With regard to affirmative action, Arcidiacono (2005) simulates how African American educational choices would change if they faced white admission and financial aid rules. Past research

[157] Recent work that specifies and estimates equilibrium models of the college market include Epple et al. (2006) and Fu (2009).

[158] Estimation is based on the EM algorithm developed in Dempster et al. (1977) and adapted to DCDP models by Arcidiacono and Jones (2003).

has shown that racial preference in the admissions process is a practice mainly at top tier institutions. Model simulations show that removing racial advantages in financial aid substantially reduces the number of African Americans who attend college and that removing advantages in admission reduces the number attending top-tier schools. However, even though such policies affect the college choice decision, they do not do much to alter lifetime earnings, which is in large part determined by initial endowments, in line with Keane and Wolpin's (1997) earlier finding.

Second, Arcidiacono (2005) uses the model to examine the reasons for large earnings and ability differences across college majors, in particular the high earnings premiums for natural science and business majors. Arcidiacono (2005) finds that monetary premia for certain majors cannot explain ability sorting across majors. Instead, almost all of the sorting occurs because of differing preferences for majors (and the jobs associated with those majors) by initial abilities. Differences in math ability is shown to be an especially important factor explaining both labor market returns and sorting across majors.

Job training

Cohen-Goldner and Eckstein (2008) uses a DCDP model to study the impact of a job training program on labor mobility and human capital accumulation. Their data consist of a short panel of observations on 419 prime age male immigrants in Israel who came from the former Soviet Union. Many of these immigrants were highly skilled upon arrival to Israel, but some of their skills were not directly transferable to the Israeli labor market. A typical pattern in the data is that immigrants start out as unemployed, move to blue collar jobs and then gradually move into white collar jobs. The government offers these immigrants a language course and job training courses to facilitate their employment transition, with a requirement that they pass a test in the Hebrew language to participate in training. One of the goals of Cohen-Goldner and Eckstein (2008) is to study the effects of these local training courses on labor market outcomes.

To this end, the paper estimates a DCDP model in which immigrants can be in one of the following states: employed in a blue-collar occupation, employed in a white-collar occupation, attending a training course in a blue-collar occupation, attending a training course in a white-collar occupation, or unemployed. An immigrant's language ability (which is self-reported) is assumed to evolve exogenously. In each period, individuals have some probability of receiving a white- or blue-collar job offer and of receiving a white-collar training offer. Blue-collar training is always an option for those with sufficient knowledge of Hebrew. Wage offers depend on the immigrants' accumulated human capital, which in turn depends on accumulated experience and training, language fluency and imported skills. The model includes four unobserved types to capture unobservable heterogeneity. It is estimated by simulated maximum likelihood.

Using the model, Cohen-Goldner and Eckstein (2008) estimate the returns from job training, controlling through the decision model for selectivity into training. They find that participating in white-collar training increases mean accepted wages by 6%

and blue-collar training by 9.8%. Participating in white-collar training also doubles the white-collar job offer probability. The net present value of government sponsored training to the immigrants is estimated to be 2.8-3.7%.

Models of early childhood investments

As earlier described, much of the work on understanding the sources of inequality in lifetime earnings and utility among individuals emphasizes the role of pre-market factors. A small literature in economics seeks to model in a dynamic setting how parental investments affect human capital formation during childhood and adolescent years and how parental investment levels are chosen. Bernal (2008) develops and estimates a DCDP model of employment and child care decisions of women, using the NLSY-Children dataset. She finds that mother's employment status and child care choices affect children's cognitive achievement outcomes, as measured by scores on the PPVT and PIAT (math and reading). Having a mother who works full-time and uses child care reduces test scores by about 2%. She uses the model to explore the effects of policy interventions that include child care subsidies and maternity leave entitlements, which she finds, on average, have adverse affects on pre-school aged children's cognitive outcomes. For example, a 35% child care subsidy, which increases the labor supply of mothers, reduces test scores by 0.23%-1.8%, depending on the age of the child.

Recent work by Cunha et al. (2010) develops new methods for estimating the so-called "technology of human capital formation." In particular, they develop and estimate a multistage, dynamic model of the evolution of cognitive and non-cognitive skills as determined by parental investments in different periods of a child's life. As in Bernal (2008), they use the NLSY-Children data and measure cognitive skills by math and reading scores. Non-cognitive skills would include such factors as motivation, self-efficacy, stubbornness and temperament, for which empirical measures are also available. The paper posits a conceptual framework in which current skills are a function of previous skill levels and intervening parental investments, parental skills, and unobserved components. It is assumed that parents sequentially choose investment levels to maximize their child's expected net present value of lifetime earnings, which implies that investment is endogenous in the estimation of the skill production technology. The paper develops an approach for addressing the endogeneity problem that jointly estimates the skill production function along with the investment policy function that can be viewed as an approximation to the decision rule from the dynamic programming problem.

A major goal of the paper is to recover substitution parameters that govern the relative important of early versus late parental investment for subsequent lifetime achievement. Cunha et al. (2010) find that investments in the early years are particularly important for the formation of adult cognitive skills and that investments in cognitive skill are much more productive if applied at younger ages. On the other hand, for non-cognitive skills, there are not dramatic differences in the productivity of investments across different life-cycle stages.

4.3.4. Summary

There is a rather consistent finding that human capital accumulation that has already occurred by the age of college attendance decisions plays a large role in subsequent labor market outcomes. Thus, policies like tuition subsidies, student loans, affirmative action, etc., aimed at affecting college attendance, may come too late in the life-cycle to remediate the main factors that lead to inequality in outcomes.[159] The literature on schooling began with and has always emphasized the college attendance decision (see Walsh, 1935). If the goal is to understand the determinants of inequality and the effects of policies for reducing inequality, then the results to date strongly suggest that the focus should shift to human capital investment decisions at earlier ages.

5. CONCLUDING REMARKS—HOW CREDIBLE ARE DCDP MODELS?

As we have illustrated, applications of the DCDP approach have addressed challenging and important questions often involving the evaluation of counterfactual scenarios or policies. The ambitiousness of the research agenda that the DCDP approach can accommodate is a major strength. This strength is purchased at a cost. To be able to perform such counterfactual analyses in such a variety of settings, DCDP models must rely on extra-theoretic modeling choices, including functional form and distributional assumptions. It is tempting to dismiss the approach for that reason, although we see no other empirical methodology with which to replace it. All approaches fall short of an assumption-free ideal that does not and is likely never to exist. And, as we have noted, DCDP researchers have taken seriously the need to provide credible validation.

There are two approaches to model validation, stemming from different epistemological perspectives. The first is the view that knowledge is absolute, that is, there exists a "true" decision-theoretic model from which observed data are generated. This leads naturally to a model validation strategy based on testing the validity of the model's behavioral implications and/or testing the fit of the model to the data. A model is deemed invalid if it is rejected according to some statistical criterion. Rejected or invalid models are discarded.

The second approach is based on a pragmatic epistemological view, in which it is acknowledged that all models are necessarily simplifications of agents' actual decision-making behavior. Hypothesis testing as a means of model validation or selection is eschewed because, given enough data, all models would be rejected as true models. In this pragmatic view, there is no true decision-theoretic model, only models that perform better or worse in addressing particular questions. Models are chosen that are "best" for some specific purpose; alternative models may be valid for different purposes.

159 The DCDP schooling models described previously take endowments at college entry ages as given. It is still an open question whether college subsidies would induce an increase in parental investments at younger ages and thus affect the endowments.

Decision-theoretic models are typically designed and estimated with the goal of predicting the impact on economic agents of changes in the economic environment. Thus, one criterion for model validation/selection that fits within the "pragmatic" view is to examine a model's predictive accuracy, namely, how successful the model is at predicting outcomes of interest within the particular context for which the model was designed. In contrast, in the absolutist view, a model would be considered useful for prediction only if it were not rejected on statistical grounds, even though non-rejection does not necessarily imply predicted effects will be close to actual effects. Nor will non-rejected models necessarily outperform rejected models in terms of their (context-specific) predictive accuracy.

Can one provide convincing evidence about the credibility of these exercises? Put differently, how can DCDP models be validated and choices be made among competing models?

There are a number of possible approaches to model validation/selection.

1. Robustness to assumptions: One method is to check how robust the empirical findings are to alternative assumptions. Although, in principle, such a method would provide evidence on the credibility of any particular set of findings, the number of assumptions in these models, their computational burden, and disagreements among researchers as to the *a priori* importance of particular assumptions, has led practitioners to limit the use of this approach.[160] That is not to say that evidence accumulated from the estimation of models by different researchers, each with different modeling inclinations, is not valuable.[161] Indeed, contradicting findings could be very revealing.

2. Within-sample model fit: DCDP papers commonly assess model fit to the estimation sample, often, but not always, using formal statistical tests. The problem with basing validation on model fit is that, like nonstructural estimation, model building is an inductive as well as deductive exercise. The final specification results from a process in which the model structure is revised as estimation proceeds, by adding parameters and changing functional forms, as deficiencies in model fit are discovered. This process of repeated model pre-testing invalidates the application of standard formal statistical tests. Nevertheless, it is interesting to note that formal tests generally reject DCDP models. Although these models tend to have a lot of parameters, sometimes numbering into the hundreds, given the extensiveness of the data moments that these models attempt to fit, the models are actually parsimonious. In practice, researchers tend to stop well short of fitting the model to idiosyncratic features of the data just to improve model fit.

160 Of course, robustness by itself cannot be conclusive; all of the models could give similarly biased results.

161 Such prejudices are revealed by the contrast between the structure of the DCDP model that Todd and Wolpin (2006) used to evaluate a conditional cash transfer program in Mexico and the model used by Attanasio et al. (2005). As another example, there are several applications of DCDP models applied to traditional topics that take a behavioral economics view. As seen, Paserman (2008) studies a job search model. In addition, Fang and Silverman (2009) study a model of women's welfare participation assuming that agents use hyperbolic discounting.

3. Out-of-sample validation: Out-of-sample validation relies on there being sample data not used in estimation, but that is assumed to come from the same underlying population. This validation sample can have a number of sources. One source for the validation sample is based on regime shifts. McFadden and Talvitie (1977), for example, estimated a random utility model (RUM) of travel demand before the introduction of the San Francisco Bay Area Rapid Transit (BART) system, obtained a forecast of the level of patronage that would ensue, and then compared the forecast to actual usage after BART's introduction. McFadden's model validation treats pre-BART observations as the estimation sample and post-BART observations as the validation sample. The validation exercise exploited data that were unavailable at the time of estimation to validate their model.

Some researchers have deliberately held out data to use for validation purposes. Lumsdaine et al. (1992), for example, estimated a model of the retirement behavior of workers in a single firm who were observed before and after the introduction of a temporary one-year pension window. They estimated several models on data before the window was introduced and compared the forecast of the impact of the pension window on retirement based on each estimated model to the actual impact as a means of model validation and selection. Keane and Moffitt (1998) estimated a model of labor supply and welfare program participation using data after federal legislation (OBRA 1981) that significantly changed the program rules. They used the model to predict behavior prior to that policy change. Keane and Wolpin (2007) estimated a model of welfare participation, schooling, labor supply, marriage and fertility on a sample of women from five US states and validated the model based on a forecast of those behaviors on a sixth state. The validation sample was purposely drawn from a state in which welfare benefits were significantly lower than in the estimation sample.

Randomized social experiments have also provided opportunities for model validation and selection. Wise (1985) exploited a housing subsidy experiment to evaluate a model of housing demand. In the experiment, families that met an income eligibility criterion were randomly assigned to control and treatment groups. The latter were offered a rent subsidy. The model was estimated using only control group data and was used to forecast the impact of the program on the treatment group. The forecast was compared to its actual impact. More recently, Todd and Wolpin (2006) used data from a large-scale school subsidy experiment in Mexico, where villages were randomly assigned to control and treatment groups. Using only the control villages, they estimated a behavioral model of parental decisions about child schooling and work, as well as family fertility. The validity of the model was then assessed according to how well it could forecast (predict) the behavior of households in the treatment villages.[162]

[162] Similarly, Lise et al. (2003) used data from a Canadian experiment designed to move people off of welfare and into work to validate a calibrated search-matching model of labor market behavior. Bajari and Hortacsu (2005) employ a similar validation methodology in the case of a laboratory auction experiment.

As should be clear from this discussion, model validation, and model building more generally, are part art and part science. For this reason, researchers will attach different priors to a model's credibility, different weights to the validation evidence, and may, therefore, come to different conclusions about the plausibility of the results. Presumably, disagreements can be reduced as confirmatory or contradictory evidence is accumulated. Whatever empirical approach to inference is adopted, structural or nonstructural, researchers should strive to provide as much validation evidence as the data and methods permit.

REFERENCES

Aguirregebaria, V., Mira, P., Dynamic discrete choice structural models: a survey. Journal of Econometrics (forthcoming).

Altonji, J., 1986. Intertemporal substitution in labor supply: evidence from micro data. Journal of Political Economy 94 (Part 2), S176–S215.

Altug, S., Miller, R.A., 1990. Household choices in equilibrium. Econometrica 58, 543–570.

Altug, S., Miller, R.A., 1998. The effect of work experience on female wages and labor supply. Review of Economic Studies 65, 45–85.

Albrecht, J., Axell, B., 1984. An equilibrium model of search employment. Journal of Political Economy 92, 824–840.

Arcidiacono, P., 2005. Affirmative action in higher education: how do admission and financial aid rules affect future earnings? Econometrica 73, 1477–1524.

Arcidiacono, P., Jones, J.B., 2003. Finite mixture distributions, sequential likelihood, and the EM algorithm. Econometrica 71, 933–946.

Arcidiacono, P., Miller, R.A., 2008. CCP estimation of dynamic discrete choice models with unobserved heterogeneity. Mimeo, Duke University.

Attanasio, O., Meghir, C., Santiago, A., 2005. Education choices in Mexico: Using a structural model and a randomized experiment to evaluate Progresa. Mimeo, University College London.

Bajari, P., Hortacsu, A., 2005. Are structural estimates of auction models reasonable? Evidence from experimental data. Journal of Political Economy 113, 703–741.

Baldwin, R.E., Cain, G.C., 2000. Shifts in relative US wages: the role of trade, technology, and factor endowments. Review of Economics and Statistics 82, 580–595.

Becker, G.S., 1964. Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education. National Bureau of Economic Research, New York.

Becker, G.S., 1967. Human capital and the personal distribution of income: an analytical approach. Woytinsky Lecture No. 1. University of Michigan, Ann Arbor, Institute of Public Administration.

Bellman, R., 1957. Dynamic Programming. Princeton University Press, Princeton.

Bellman, R., Kalaba, R., Kootkin, B., 1963. Polynomial approximation—a new computational technique in dynamic programming allocation processes. Mathematics of Computation 1, 155–161.

Belzil, C., Hansen, J., 2002. Unobserved ability and the return to schooling. Econometrica 70, 2075–2091.

Ben-Porath, Y., 1967. The production of human capital and the life-cycle of earnings. Journal of Political Economy 75, 352–365.

Bernal, R., 2008. The effect of maternal employment and child care on children's cognitive development. International Economic Review 49, 1173–1209.

Blau, D, 1991. Search for nonwage job characteristics: a test of the reservation wage hypothesis. Journal of Labor Economics 9, 186–205.

Blundell, R.W., Walker, I., 1986. A life cycle consistent empirical model of labor supply using cross section data. Review of Economic Studies 53, 539–558.

Blundell, R.W., Duncan, A., Meghir, C., 1998. Estimating labour supply responses using tax policy reforms. Econometrica 66, 827–861.

Bontemps, C., Robin, J.-M., van den Berg, G.J., 1999. An empirical equilibrium job search model with search on the job and heterogeneous workers and firms. International Economic Review 40, 1039–1074.
Bontemps, C., Robin, J.-M., van den Berg, G.J., 2000. Equilibrium search with continuous productivity dispersion: theory and nonparametric estimation. International Economic Review 41, 305–358.
Bound, J., Johnson, G., 1992. Changes in the structure of wages in the 1980s: an evaluation of alternative explanations. American Economic Review 82, 371–392.
Burdett, K., 1978. A theory of employee job search and quit rates. American Economic Review 68, 212–220.
Burdett, K., Mortensen, D.T., 1998. Wage differentials, employer size and unemployment. International Economic Review 39, 257–273.
Burdett, K., Ondrich, J.I., 1985. How changes in labor demand affect unemployed workers. Journal of labor Economics 3, 1–10.
Cahuc, P., Postel-Vinay, F., Robin, J.-M., 2006. Wage bargaining with on the job search: theory and evidence. Econometrica 74, 323–364.
Cameron, S.V., Heckman, J.J., 1998. Life cycle schooling and dynamic selection bias: models and evidence for five cohorts of American males. Journal of Political Economy 106, 262–333.
Cameron, S.V., Heckman, J.J., 1999. Can tuition policy combat rising wage inequality? In: Kosters, M. (Ed.), Financing College Tuition: Government Policies Social Priorities. AEI Press, Washington, DC, pp. 76–121.
Card, D., 2001. Estimating the returns to schooling: progress on some persistent econometric problems. Econometrica 69, 1127–1160.
Chamberlain, G., 1984. Panel data. In: Griliches, Z., Intriligator, M. (Eds.), Handbook of Econometrics, vol. 2. North Holland, Amsterdam.
Ching, A., Imai, S., Ishihara, M., Jain, N., 2010. A guide to Bayesian estimation of discrete choice models with an application to a store level reward program. Mimeo. University of Toronto.
Christensen, B.J., Kiefer, N.M., 1991. The exact likelihood function for an empirical job search model. Econometric Theory 7, 464–486.
Christensen, B.J., Kiefer, N.M., 2009. Economic Modeling and Inference. Princeton University Press, Princeton.
Classen, K.P., 1977. The effect of unemployment insurance on the duration of unemployment and subsequent earnings. Industrial and Labor Relations Review 30, 438–444.
Cogan, J., 1981. Fixed costs and labor supply. Econometrica 49, 945–964.
Cohen-Goldner, S., Eckstein, Z., 2008. Labor mobility of immigrants: training, experience, language and opportunities. International Economic Review 49, 837–874.
Cunha, F., Heckman, J.J., Schennach, S.M., 2010. Estimating the technology of cognitive and noncognitive skill formation. Econometrica 78, 883–931.
Daula, T., Moffitt, R., 1995. Estimating dynamic models of quit behavior: the case of military reenlistment. Journal of Labor Economics 13, 499–523.
Danforth, J.P., 1979. On the role of consumption and decreasing absolute risk aversion in the theory of job search. In: Lippmand, S.A., McCall, J.J. (Eds.), Studies in the Economics of Search. North Holland, New York, pp. 109–131.
Della Vigna, S., Paserman, M.D., 2005. Job search and impatience. Journal of Labor Economics 23, 527–588.
Dempster, A.P., Laird, M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society B 39, 1–38.
Diamond, P.A., 1970. A model of price adjustment. Journal of Economic Theory 3, 156–168.
Diamond, P.A., Maskin, E., 1979. An equilibrium analysis of search and breach of contracts I: steady states. Bell Journal of Economics 10, 282–316.
Dominitz, J., Manski, C.F., 1996. Eliciting student expectations of the returns to schooling. Journal of Human Resources 31, 1–26.
Dominitz, J., Manski, C.F., 1997. Using expectations data to study subjective income expectations. Journal of the American Statistical Association 92, 855–867.
Eckstein, Z., Nagypal, E., 2004. The evolution of US earnings inequality: 1961-2002. Federal Reserve Bank of Minneapolis Quarterly Review 28, 10–29.
Eckstein, Z., Vandenberg, G., 2007. Empirical labor search: a survey. Journal of Econometrics 136, 531–564.

Eckstein, Z., Wolpin, K.I., 1989a. The specification and estimation of dynamic stochastic discrete choice models. Journal of Human Resources 24, 562–598.
Eckstein, Z., Wolpin, K.I., 1989b. Dynamic labor force participation of married women and endogenous wage growth. Review of Economic Studies 56, 375–390.
Eckstein, Z., Wolpin, K.I., 1990. Estimating a market equilibrium search model from panel data on individuals. Econometrica 58, 783–808.
Eckstein, Z., Wolpin, K.I., 1995. Duration to first job and the return to schooling: estimates from a search-matching model. Review of Economic Studies 62, 263–286.
Eckstein, Z., Wolpin, K.I., 1999. Why youths drop out of high school: the impact of preferences, opportunities, and abilities. Econometrica 67, 1295–1339.
Epple, D., Romano, R., Sieg, H., 2006. Admission, tuition, and financial aid policies in the market for higher education. Econometrica 74, 885–928.
Fang, H., Silverman, D., 2009. Time inconsistency and welfare program participation: evidence from the NLSY. International Economic Review 50, 1043–1078.
Ferrall, C., 1997. Unemployment insurance eligibility and the school to work transition in Canada and the United States. Journal of Business Economics and Statistics 15, 115–129.
Flinn, C.J., 2006. Minimum wage effects on labor market outcomes under search, matching and endogenous contact rates. Econometrica 74, 1013–1062.
Flinn, C.J., Heckman, J.J., 1982. New methods for analyzing structural models of labor force dynamics. Journal of Econometrics 18, 114–142.
Francesconi, M., 2002. A joint dynamic model of fertility and work of married women. Journal of labor Economics 20, 336–380.
French, E., Taber, C., 2011. Identification of models of the labor market. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 4A. Elsevier, Amsterdam, pp. 537–617.
Fu, Chao, 2009. An equilibrium model of the college market: an empirical investigation of tuition, applications, admissions and enrollment. Mimeo, University of Pennsylvania.
Fuchs, V.R., 1980. Economic Growth and the Rise of the Service Sector. Working paper no. 386, National Bureau of Economic Research.
Gemici, A., 2007. Family migration and labor market outcomes. Mimeo, University of Pennsylvania.
Geweke, J., Keane, M.P., 2000. In: Mariano, Schuermann (Eds.), Bayesian Inference for Dynamic Discrete Choice Models without the Need for Dynamic Programming. Cambridge University Press, Cambridge.
Geweke, J., Keane, M.P., 2001. Computationally intensive methods for integration in econometrics. In: Heckman, J.J., Leamer, E.E. (Eds.), Handbook of Econometrics, vol. 5. Elsevier, Amsterdam.
Gottschalf, P., Moffitt, R., 1994. The growth of earnings instability in the US labor market. Brookings Papers on Economic Activity 2, 217–272.
Gotz, G.A., McCall, J.J., 1984. A dynamic retention model for air force officers: theory and estimates. RAND, R-3028-AF.
Gould, Eric, 2008. Marriage and career: the dynamic decisions of young men. Journal of Human Capital 2, 337–378.
Greenberger, E., Steinberg, L., 1986. When Teenagers Work: The Psychological and Social Costs of Adolescent Employment. Basic Books, New York.
Griliches, Z., 1977. Estimating the returns to schooling: some persistent econometric problems. Econometrica 45, 1–22.
Heckman, J.J., 1974. Shadow prices, market wages and labor supply. Econometrica 42, 679–694.
Heckman, J.J., 1979. Sample selection bias as a specification error. Econometrica 47, 153–162.
Heckman, J.J., 1981. The incidental parameters problem and the problem of initial conditions in estimating a discrete time—discrete data stochastic process. In: Manski, C.F., McFadden, D. (Eds.), Structural Analysis of Discrete Data with Econometric Applications. MIT Press, Cambridge MA.
Heckman, J.J., Honore, B.E., 1990. The empirical content of the Roy model. Econometrica 58, 1121–1149.
Heckman, J.J., Lochner, L., Taber, C., 1998. Explaining rising wage inequality: explorations with a dynamic general equilibrium model of earnings with heterogeneous agents. Review of Economic Dynamics 1, 1–58.
Heckman, J.J., MaCurdy, T.E., 1980. A life cycle model of female labor supply. Review of Economic Studies 47, 47–74.

Heckman, J.J., MaCurdy, T.E., 1982. Corrigendum on a life cycle model of female labor supply. Review of Economic Studies 49, 659–660.
Heckman, J.J., Polachek, S., 1974. Empirical evidence on the functional form of the earnings-schooling relationship. Journal of the American Statistical association 69, 350–354.
Heckman, J.J., Sedlacek, G., 1985. Heterogeneity, aggregation and market wage functions: an empirical model of self-selection in the labor market. Journal of Political Economy 93, 1077–1125.
Heckman, J.J., Singer, B., 1984. A method for minimizing the impact of distributional assumptions in econometric models for duration data. Econometrica 52, 271–320.
Heckman, J.J., Willis, R.J., 1977. A Beta-logistic model for the analysis of sequential labor force participation by married women. Journal of Political Economy 85, 27–58.
Hornstein, A., Krusell, P., Violante, G.L., 2005. The effects of technical change on labor market inequalities. In: Aghion, P., Durlauf, S. (Eds.), Handbook of Economic Growth. Elsevier, Amsterdam.
Hotz, J., Miller, R.A., 1993. Conditional choice probabilities and the estimation of dynamic models. Review of Economic Studies 60, 497–530.
Hotz, J., Miller, R.A., Sanders, S., Smith, J., 1994. A simulation estimator for dynamic models of discrete choice. Review of Economic Studies 61, 265–289.
Ichimura, H., Taber, C., 2002. Semi-parametric reduced form estimation of tuition subsidies. American Economic Association, Papers and Proceedings 286–292.
Imai, S., Jain, N., Ching, A., 2009. Bayesian estimation of discrete choice models. Econometrica 77, 1865–1900.
Imai, S., Keane, M.P., 2004. Intertemporal labor supply and human capital accumulation. International Economic Review 45, 601–641.
Jovanovic, B., 1979. Job matching and the theory of turnover. Journal of Political Economy 87, 972–990.
Juhn, C., Murphy, K.M., Pierce, B., 1993. Wage inequality and the rise in returns to skill. Journal of Political Economy 101, 410–442.
Kane, T., 1999. The Price of Admission: Rethinking How Americans Pay for College. The Brookings Institution, Washington, DC.
Katz, L.F., Autor, D.H., 1999. Changes in the age structure and earnings inequality. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 3. North Holland, Amsterdam, pp. 1463–1555.
Katz, L.F., Murphy, K.M., 1992. Changes in relative wages, 1963-1987: supply and demand factors. Quarterly Journal of Economics 107, 35–78.
Keane, M.P., 2009a. Labor supply and taxes: a survey. Mimeo. University of Technology Sidney.
Keane, M.P., 2009b. Income taxation in a life cycle model with human capital. Mimeo. University of Technology Sidney.
Keane, M.P., Moffitt, R., 1998. A structural model of multiple welfare program participation and labor supply. International Economic Review 39, 553–590.
Keane, M.P., Moffitt, R., Runkle, D., 1988. Real wages over the business cycle: estimating the impact of heterogeneity with micro data. Journal of Political Economy 96, 1232–1266.
Keane, M.P., Sauer, R.M., 2009. A computationally practical simulation estimator for panel data models with unobserved endogenous state variables. International Economic Review (forthcoming).
Keane, M.P., Wolpin, K.I., 1994. The solution and estimation of discrete choice dynamic programming models by simulation: Monte Carlo evidence. Review of Economics and Statistics 648–672.
Keane, M.P., Wolpin, K.I., 1997. The career decisions of young men. Journal of Political Economy 105, 473–522.
Keane, M.P., Wolpin, K.I., 2000. Eliminating race differences in school attainment and labor market success. Journal of Labor Economics 18, 614–652.
Keane, M.P., Wolpin, K.I., 2001. The effect of parental transfers and borrowing constraints on educational attainment. International Economic Review 42, 1051–1103.
Keane, M.P., Wolpin, K.I., 2002. Estimating welfare effects consistent with forward looking behavior: an article in two parts. Journal of Human Resources 37, 570–622.
Keane, M.P., Wolpin, K.I., 2007. Exploring the usefulness of a non-random holdout sample for model validation: welfare effects on female behavior. International Economic Review 48, 1351–1378.
Keane, M.P., Wolpin, K.I., 2009. Empirical applications of discrete choice dynamic programming models. Review of Economic Dynamics 12, 1–22.

Keane, M.P., Wolpin, K.I., 2010. The role of labor and marriage markets, preference heterogeneity, and the welfare system in the life cycle decisions of black, Hispanic and white women. International Economic Review 51, 851–892.
Kiefer, N.M., Neumann, G.R., 1993. Wage dispersion and homogeneity: the empirical equilibrium search model. In: Bunzel, H., et al. (Eds.), Panel Data and Labor Market Dynamics. North Holland, Amsterdam.
Kimmel, J., Knieser, T.J., 1998. New evidence on labor supply: employment vs. hours elasticities by sex and marital status. Journal of Monetary Economics 42, 289–301.
Krusell, P., Ohanian, L.E., Ríos-Rull, J.V., Violante, G.L., 2000. Capital-skill complementarity and inequality: A macroeconomic analysis. Econometrica 68, 1029–1053.
Lee, D., 2005. An estimable dynamic general equilibrium model of school, work and occupational choice. International Economic Review 46, 1–34.
Lee, D., Wolpin, K.I., 2006. Intersectoral labor mobility and the growth of the service sector. Econometrica 74, 1–46.
Lee, D., Wolpin, K.I., 2010. Accounting for wage and employment changes in the US from 1968-2000: a dynamic model of labor market equilibrium. Journal of Econometrics 156, 68–85.
Lentz, R., 2009. Optimal unemployment insurance in an estimated job search model with savings. Review of Economic Studies 12, 37–57.
Lise, J., Seitz, S., Smith, J., 2003. Equilibrium policy experiments and the evaluation of social programs. Mimeo, Queens University.
Lumsdaine, Robin L., Stock, James H., Wise, David A., 1992. Pension plan provisions and retirement: men and women, medicare, and models. In: Wise, D.A. (Ed.), Studies in the Economics of Aging, University of Chicago Press, Chicago.
McCall, J.J., 1970. Economics of information and job search. Quarterly Journal of Economics 84, 113–126.
MaCurdy, T.E., 1981. An empirical model of labor supply in a life cycle setting. Journal of Political Economy 89, 1059–1085.
MaCurdy, T.E., 1983. A simple scheme for estimating an intertemporal model of labor supply and consumption in the presence of taxes and uncertainty. International Economic Review 24, 265–289.
McFadden, D., 1989. A method of simulated moments for estimation of discrete response models with numerical integration. Econometrica 57, 995–1028.
McFadden, Daniel, Talvitie, A.P., associates, 1977. Validation of disaggregate travel demand models: Some tests. Urban demand forecasting project, final report. Volume V, Institute of Transportation Studies, University of California, Berkeley.
Manski, C.F., 2004. Measuring expectations. Econometrica 72, 1329–1376.
Matzkin, R.L., 1993. Nonparametric and distribution-free estimation of the binary threshhold crossing and the binary choice models. Econometrica 60, 239–270.
Mazzocco, M., Yamaguchi, S., 2006. Labor supply, wealth dynamics and marriage decisions. Mimeo. UCLA.
Meyer, B., 1990. Unemployment insurance and unemployment spells. Econometrica 58, 757–782.
Miller, R.A., 1984. Job matching and occupational choice. Journal of Political Economy 92, 1086–1120.
Miller, R.A., 1997. Estimating models of dynamic optimization with microeconomic data. In: Pesaran, M.H., Schmidt, P. (Eds.), Handbook of Applied Econometrics: Microeconomics. Basil Blackwell.
Mincer, J., 1958. Investment in human capital and personal income distribution. Journal of Political Economy 66, 281–302.
Mincer, J., 1962. Labor force participation of married women: a study of labor supply. In: Aspects of Labor Economics. National Bureau of Economic Research, Princeton, NJ, pp. 63–97.
Moffitt, R., 1983. An economic model of welfare stigma. American Economic Review 73, 1023–1035.
Mortensen, D.T., 1970. A theory of wage and employment dynamics. In: Phelps, E.S., et al. (Eds.), Microeconomic Foundations of Employment and Inflations Theory. W.W. Horton, New York.
Mortensen, D.T., 1976. Unemployment insurance and job search decisions. Industrial and Labor Relations Review 30, 505–517.
Mortensen, D.T., 1982. The matching process as a noncooperative game. In: McCall, J.J. (Ed.), The Economics of Information and Uncertainty. NBER, University of Chicago Press, Chicago.
Mortensen, D.T., 1986. Job search and labor market analysis. In: Handbook of Labor Economics. North-Holland, Amsterdam.

Murphy, K.M., Welch, F., 1992. The structure of wages. Quarterly Journal of Economics 107, 285–326.
Murphy, K.M., Welch, F., 1993. Occupational change and the demand for skill: 1940-1990. American Economic Review 83, 122–126.
Norets, A., 2009. Inference in dynamic discrete choice models with serially correlated unobserved state variables. Econometrica 77, 1665–1682.
Pagan, A., Ullah, A., 1999. Nonparametric Econometrics. Cambridge University Press, Cambridge.
Pakes, A., 1986. Patents as options: some estimates of the value of holding European patent stocks. Econometrica 54, 755–784.
Paserman, M.D., 2008. Job search and hyperbolic discounting: structural estimation and policy evaluation. The Economic Journal 118, 1418–1452.
Postel-Vinay, F., Robin, J.-M., 2002. Equilibrium wage dispersion and with heterogeneous workers and firms. Econometrica 70, 2295–2350.
Rendon, S., 2006. Job search and asset accumulation under borrowing constraints. International Economic Review 47, 233–264.
Rosen, S., 1977. Human capital: a survey of empirical research. In: Ehrenberg, R. (Ed.), Research in Labor Economics, vol. 1. JAI Press, Greenwhich, CT.
Roy, A., 1951. Some thoughts on the distribution of earnings. Oxford Economic Papers 3, 135–146.
Rust, J., 1987. Optimal replacement of GMC bus engines: an empirical model of Harold Zurcher. Econometrica 55, 999–1034.
Rust, J., 1993. Structural estimation of markov decision models. In: Engle, R., McFadden, D. (Eds.), Handbook of Econometrics, vol. 4. North Holland, Amsterdam.
Rust, J., 1994. Numerical dynamic programming in economics. In: Amman, H.M., et al. (Eds.), Handbook of Computational Economics. North Holland, Amsterdam.
Rust, J., 1997. Using randomization to break the curse of dimensionality. Econometrica 65, 487–516.
Sauer, R., 1998. Job mobility and the market for lawyers. Journal of Political Economy 106, 147–171.
Sargent, T.J., 1987. Dynamic Macroeconomic Theory. Harvard University Press, Cambridge, MA.
Sauer, R., 2004. Education financing and lifetime earnings. Review of Economic Studies 71, 1189–1216.
Shaw, K., 1989. Life cycle labor supply with human capital accumulation. International Economic Review 30, 431–456.
Stigler, G., 1961. The economics of information. Journal of Political Economy 69, 213–225.
Stinebrickner, T.R., 2000. Serially correlated variables in dynamic, discrete choice models. Journal of Applied Econometrics 15, 595–624.
Stinebrickner, T.R., 2001. A dynamic model of teacher labor supply. Journal of Labor Economics 19, 196–230.
Stern, S., 1989. Estimating a simultaneous search model. Journal of Labor Economics 7, 348–369.
Sullivan, P., 2010. A dynamic analysis of educational attainment, occupational choices and job search. International Economics Review 51, 289–317.
Tartari, M., 2007. Divorce and Cognitive Achievement of Children. Mimeo. Yale University.
Todd, P., Wolpin, K.I., 2006. Assessing the impact of a school subsidy program in Mexico: using a social experiment to validate a dynamic behavioral model of child schooling and fertility. American Economic Review 96, 1384–1417.
Todd, P., Wolpin, K.I., 2010. Ex-ante evaluation of social programs. Annals of Economics and Statistics.
Van den Berg, G.J., 1990. Nonstationarity in job search theory. Review of Economic Studies 57, 255–277.
Van den Berg, G.J., Ridder, G., 1998. An empirical equilibrium search model of the labor market. Econometrica 66, 1183–1221.
Van der Klaauw, W., 1996. Female labour supply and marital status decisions: a life cycle model. Review of Economic Studies 63, 199–235.
Van der Klaauw, W., 2000. On the use of expectations data in estimating structural dynamic models: an analysis of career choices. Mimeo. University of North Carolina.
Van der Klaauw, W., Wolpin, K.I., 2008. Social security and the retirement and savings behavior of low income households. Journal of Econometrics 145, 21–42.
Van Soest, Woittiez, A.I., Kapteyn, A., 1990. Labor supply, income taxes and hours restrictions in the Netherlands. Journal of Human Resources 25, 517–558.

Weiss, Y., Gronau, R., 1981. Expected interruptions in labour force participation and sex-related differences in earnings. Review of Economic Studies 48, 607–619.
Walsh, J.R., 1935. Capital concept applied to man. Quarterly Journal of Economics 49, 255–285.
Welch, F., 1977. What have we learned from empirical studies of unemployment insurance? Industrial and Labor Relations Review 30, 450–461.
Welch, F., 2000. Growth in womens' relative wages and in inequality among men:one phenomenon or two. American Economic Review 90, 444–449.
Willis, R., Rosen, S., 1979. Education and self-selection. Journal of Political Economy 87, S7–S36.
Wise, D., 1985. A behavioral model versus experimentation: the effects of housing subsidies on rent. In: Brucker, P., Pauly, R. (Eds.), Methods of Operations Research, 50. pp. 441–489.
Wolinsky, A., 1987. Matching, search and bargaining. Journal of Economic Theory 42, 311–333.
Wolpin, K.I., 1984. A dynamic stochastic model of fertility and child mortality. Journal of Political Economy 92, 852–874.
Wolpin, K.I., 1987. Estimating a structural search model: the transition from school to work. Econometrica 55, 801–817.
Wolpin, K.I., 1992. The determinants of black-white differences in early employment careers: search, layoffs, quits and endogenous wage growth. Journal of Political Economy 100, 535–560.
Wolpin, K.I., 1995. Empirical Methods for the Study of Labor Force Dynamics. Harwood Academic Publishers, Luxembourg.
Wolpin, K.I., 2003. Wage Equations and Education Policy. In: Drewatripont,, et al. (Eds.), Advances in Econometrics and Economics, vol. II. Cambridge University Press, Cambridge.

This page intentionally left blank

CHAPTER 5

Program Evaluation and Research Designs☆

John DiNardo*, David S. Lee**

* University of Michigan and NBER

** Princeton University and NBER

Contents

☆ We are grateful to Diane Alexander and Pauline Leung, who provided outstanding research assistance. We thank Orley Ashenfelter, David Card, Damon Clark, Nicole Fortin, Thomas Lemieux, Enrico Moretti, Phil Oreopolous, Zhuan Pei, Chris Taber, Petra Todd, John Van Reenen, and Ken Wolpin for helpful suggestions, comments, and discussions.

Handbook of Labor Economics, Volume 4a
ISSN 0169-7218, DOI 10.1016/S0169-7218(11)00411-4

Abstract

This chapter provides a selective review of some contemporary approaches to program evaluation. One motivation for our review is the recent emergence and increasing use of a particular kind of "program" in applied microeconomic research, the so-called Regression Discontinuity (RD) Design of Thistlethwaite and Campbell (1960). We organize our discussion of these various research designs by how they secure internal validity: in this view, the RD design can been seen as a close "cousin" of the randomized experiment. An important distinction which emerges from our discussion of "heterogeneous treatment effects" is between ex post (descriptive) and ex ante (predictive) evaluations; these two types of evaluations have distinct, but complementary goals. A second important distinction we make is between statistical statements that are descriptions of our knowledge of the program assignment process and statistical statements that are structural assumptions about individual behavior. Using these distinctions, we examine some commonly employed evaluation strategies, and assess them with a common set of criteria for "internal validity", the foremost goal of an ex post evaluation. In some cases, we also provide some concrete illustrations of how internally valid causal estimates can be supplemented with specific structural assumptions to address "external validity": the estimate from an internally valid "experimental" estimate can be viewed as a "leading term" in an extrapolation for a parameter of interest in an ex ante evaluation.

1. INTRODUCTION

This chapter provides a selective review of some contemporary approaches to program evaluation. Our review is primarily motivated by the recent emergence and increasing use of the a particular kind of "program" in applied microeconomic research, the so-called Regression Discontinuity (RD) Design of Thistlethwaite and Campbell (1960). In a recent survey, Lee and Lemieux (2009) point out that the RD design has found good use in a wide variety of contexts, and that over the past decade, the way in which researchers view the approach has evolved to a point where it is now considered to yield highly credible and transparent causal inferences. At the time of the last volumes of the Handbook of Labor Economics, the RD design was viewed simultaneously as a "special case" of Instrumental Variables (IV) (Angrist and Krueger, 1999) and a "special case" of a "selection on observables", or matching approach (Heckman et al., 1999). Recent theoretical analyses and the way in which practitioners interpret RD designs reveal a different view; Lee and Lemieux (2009) point out that the RD design can be viewed as a close "cousin" of the randomized experiment. In this chapter, we provide an extended discussion of this view, and also discuss some of the issues that arise in the practical implementation of the RD design. The view of the RD design as a "cousin" of the randomized experiment leads to our second, broader objective in this review: to chart out this perspective's implicit "family tree" of commonly used program evaluation approaches.[1]

[1] Other recent reviews of common evaluation approaches include, for example, Heckman and Vytlacil (2007a,b) and Abbring and Heckman (2007).

Our discussion necessarily involves a discussion of "heterogeneous treatment effects", which is one of the central issues in a wider debate about the relative merits of "structural" versus "design-based"/"experimentalist" approaches.[2] In setting forth a particular family tree, we make no attempt to make explicit or implicit judgments about what is a "better" or "more informative" approach to conducting research. Instead, we make two distinctions that we think are helpful in our review.

First, we make a clear distinction between two very different kinds of evaluation problems. One is what could be called the *ex-post evaluation problem*, where the main goal is to document "what happened" when a particular program was implemented. The problem begins with an explicit understanding that a very particular program was run, individuals were assigned to, or self-selected into, program status in a very particular way (and we as researchers may or may not know very much about the process), and that because of the way the program was implemented, it may only be possible to identify effects for certain sub-populations. In this sense, the data and the context (the particular program) define and set limits on the causal inferences that are possible. Achieving a high degree of internal validity (a high degree of confidence that what is measured indeed represents a causal phenomenon) is the primary goal of the ex post evaluation problem.

The other evaluation problem is the *ex-ante evaluation problem*, which begins with an explicit understanding that the program that was actually run may not be the one that corresponds to a particular policy of interest. Here, the goal is not descriptive, but is instead predictive. What would be the impact if we expanded eligibility of the program? What would the effects be of a similar program if it were run at a national (as opposed to a local) level? Or if it were run today (as opposed to 20 years ago)? It is essentially a problem of forecasting or extrapolating, with the goal of achieving a high degree of external validity.[3]

We recognize that in reality, no researcher will only pursue (explicitly or implicitly) one of these goals to the exclusion of the other. After all, presumably we are interested in studying the effects of a particular program that occurred in the past because we think it has predictive value for policy decisions in the here and now. Likewise, a forecasting exercise usually begins with some assessment of how well methods perform "in-sample". Nevertheless, keeping the "intermediate goals" separate allows us to discuss more clearly how to achieve those goals, without having to discuss which of them is "more important" or ambitious, or more worthy of a researcher's attention.

The second distinction we make—and one that can be more helpful than one between "structural" and "design-based" approaches—is the one between "structural"

[2] A sampling of papers that reflects this debate would include Heckman and Vytlacil (2005), Heckman et al. (2006), Deaton (2008), Imbens (2009), Keane (2009) and Angrist and Pischke (2010).

[3] In our chapter, we will say nothing about another kind of ex ante evaluation question: what would be the effects of a program that was never run in the first place, or of a qualitatively different kind of program? See the discussion in Todd and Wolpin (2006).

and "design-based" *statistical conditions*. When we have some institutional knowledge about the process by which treatment was assigned, and when there can be common agreement about how to represent that knowledge as a statistical statement, we will label that a "D"-condition; "D" for "data-based", "design-driven", or "descriptive". These conditions are better thought of as *descriptions* of what actually generated the data, rather than *assumptions*. By contrast, when important features of the data generating process are unknown, we will have to invoke some conjectures about behavior (perhaps motivated by a particular economic model), or other aspects about the environment. When we do not literally know if the conditions actually hold, but nevertheless need them to make inferences, we will label them "S"-conditions; "S" for "structural", "subjective", or "speculative". As we shall see, inference about program effects will frequently involve a combination of "D" and "S" conditions: it is useful to be able to distinguish between conditions whose validity is secure and those conditions whose validity is not secure.

Note that although we may not know whether "S"-conditions are literally true, sometimes they will generate strong testable implications, and sometimes they will not. And even if there is a strong link between what we know about program assignment and a "D"-condition, a skeptic may prefer to treat those conditions as hypotheses; so we will also consider the testable implications that various "D"-conditions generate.

Using these distinctions, we examine some commonly employed evaluation strategies, and assess them against a common set of criteria for "internal validity". We also provide a few concrete illustrations of how the goal of an ex post evaluation are quite complementary to that of an ex ante evaluation. Specifically, for a number of the designs, where "external validity" is an issue, we show some examples where internally valid causal estimates—supplemented with specific "S"-conditions—can be viewed as a "leading term" in an extrapolation for a parameter of interest from an ex ante evaluation standpoint.

Our review of commonly employed evaluation strategies will highlight and emphasize the following ideas, some of which have long been known and understood, others that have gained much attention in the recent literature, and others that have been known for some time but perhaps have been under-appreciated:

- From an ex post evaluation standpoint, a carefully planned experiment using random assignment of program status represents the ideal scenario, delivering highly credible causal inferences. But from an ex ante evaluation standpoint, the causal inferences from a randomized experiment may be a poor forecast of what were to happen if the program where to be "scaled up". We provide a simple illustration of how this policy parameter of interest might be linked to the parameter identified from an experiment.
- When program status is described as random assignment with imperfect (and non-random) compliance, the IV (i.e. Wald) estimand delivers an average causal effect that may well not be as "local" as the usual Local Average Treatment Effect (LATE) interpretation suggests. Although LATE has been defined as the "average treatment

effect [only] for individuals whose treatment status is influenced by changing an exogenous regressor" (Imbens and Angrist, 1994), we show that a "probabilistic monotonicity" condition allows the IV estimand to be interpreted as a weighted average effect for all individuals, where the weights are proportional to the effect of the instrument on the *probability* of treatment receipt.

- From an ex post evaluation standpoint, when program status is characterized as random assignment with imperfect compliance, LATE represents "what is possible" to identify with minimal assumptions. But from an ex ante evaluation standpoint, it may not be adequate for predicting, for example, the impact of the program if receipt was mandated (the Average Treatment Effect (ATE)). We highlight the well-known fact that ATE can be viewed as an extrapolation that has LATE as its "leading term".
- Curiously, our literature search revealed that applied researchers typically do not conduct or report such extrapolations, even though the parameters of that extrapolation are identified from the same data used to compute estimates of LATE. We apply such an extrapolation for a small sampling of studies in the literature to show the differences between LATE and (one estimate of) ATE in practice.
- The presence of "local random assignment" around the threshold in a Regression Discontinuity design is not merely a "maintained" assumption, but rather a *consequence* of a structural assumption (with strong testable implications) about the extent to which agents can precisely manipulate the assignment variable.
- The discontinuity in the RD estimand generally has a less "local" interpretation than "the average effect for those individuals at the threshold". It can be viewed as a weighted average effect, where the weights are proportional to the ex ante likelihood that the value of the individual's assignment variable would lie in a neighborhood of the threshold.
- It is clear that *any* program evaluation method ultimately requires unobservables to be independent with either an instrument or treatment status itself. But there is an important difference between *assuming* that unobservables are independent of instruments or program status, and when such a condition holds as a *consequence* of a particular data generating process.
- When employing matching estimators in a "selection on observables" approach in non-experimental settings, "adding more controls" in the analysis carries a great risk of *exacerbating* any possible selection biases.

The chapter is organized as follows: in Section 2 we provide some background for our review, including our criteria for assessing various research designs; we also make some important distinctions between types of "program evaluation" that will be useful in what follows. One important distinction will be between research designs where the investigator has detailed institutional knowledge of the process by which individuals were assigned to treatment ("dominated by knowledge of the assignment process") and those research designs where such information is lacking—what we describe as

being "dominated by self-selection." In Section 3, we discuss the former: this includes both randomized controlled trials and the regression discontinuity design. In Section 4, we discuss the latter: this includes "differences-in-differences", instrumental variables ("selection on unobservables"), matching estimators, ("selection on observables"). Section 5 concludes.

2. SCOPE AND BACKGROUND

The term "program evaluation" is frequently used to describe any systematic attempt to collect and analyze information about the implementation and outcomes of a "program"—a set of policies and procedures. Although program evaluations often include "qualitative" information, such as narrative descriptions about aspects of the program's implementation, our focus will be solely on statistical and econometric evaluation. For our purposes, a program is a set of interventions, actions or "treatments" (typically binary), which are assigned to participants and are suspected of having some consequences on the outcomes experienced by the participants. Individuals who are "assigned" or "exposed" to treatment may or may not take up the treatment; when some individuals are assigned to, but do not take up the treatment we will often find it convenient to evaluate the effect of the offer of treatment (an "intent to treat analysis"), rather than the effect of the treatment *per se*, although we will examine what inferences can be made about the effect of the treatment in these situations. The problem will be to study the causal effect of the treatment when "the effects under investigation tend to be masked by fluctuations outside the experimenter's control" (Cox, 1958). Examples of programs and treatments include not only explicit social experiments such as those involving the provision of job training to individuals under the Job Training Partnership Act (JTPA) (Guttman, 1983), but also "treatments" provided outside the context of specifically designed social experiments. Some examples of the latter include the provision of collective bargaining rights to workers at firms (DiNardo and Lee, 2004), the effects of social insurance on labor market outcomes (Lemieux and Milligan, 2008), health insurance (Card et al., 2009b,a) and schooling to mothers (McCrary and Royer, 2010).

Our review will be selective. In particular, we will focus most of our attention on situations in which "institutional knowledge of the data generation process" strongly informs the statistical and econometric analysis.[4] With such a focus, a discussion of randomized controlled trials (RCTs) and the regression discontinuity design (RDD) are featured not because they are "best" in some single index ranking of "relevance", but because they often provide situations where a "tight link" between the posited statistical model and the institutional details of the experiment lends credibility to the conclusions. The statistical model employed to analyze a simple, well-designed RCT often bears a

[4] For a comprehensive discussion and review of many of these issues see the reviews of Heckman and Vytlacil (2007a,b) and Abbring and Heckman (2007).

tighter resemblance to the institutional details of the designed experiment than does, for example, a Mincerian wage regression. In this latter case, the credibility of the exercise does not rest on the fact that wages are set in the market place as a linear combination of a non-stochastic relationship between potential experience, schooling, etc. and a stochastic error term: the credibility of such an exercise instead rests on factors *other* than its close resemblance to the institutional realities of wage setting.

The distinction between these situations has sometimes been blurred: the Neyman-Holland-Rubin Model (Splawa-Neyman et al., 1990, 1935; Rubin, 1990, 1974, 1986; Holland, 1986), which we discuss later, has been used in situations both where the investigator *does* have detailed institutional knowledge of the data generating process and where the investigator *does not*. Our focus is on "the experiment that happened" rather than the "experiment we would most like to have been conducted". As others have noted, this focus can be limiting, and a given experiment may provide only limited information (if any) on structural parameters interesting to some economists (see for example Heckman and Vytlacil (2007a)). If a designed experiment assigns a package of both "remedial education" and "job search assistance" to treated individuals, for example, we may not be able to disentangle the separate effects of each component on subsequent employment outcomes. We may be able to do better if the experiment provides random assignment of each of the components separately and together, but this will depend crucially on the experiment that was actually conducted.

In adopting such a focus, we do not mean to suggest that the types of research designs we discuss should be the only ones pursued by economists and we wish to take no position on where the "marginal research dollar" should be spent or the appropriate amount of energy which should be dedicated to "structural analyses"; for some examples of some recent contributions to this debate see Deaton (2008), Heckman and Urzua (2009), Imbens (2009), Keane (2009) and Rust (2009). Moreover, even with this narrow focus there are several important subjects we will not cover, such as those involving a continuously distributed randomized instrument as in Heckman and Vytlacil (2001a); some of these issues are treated in Taber and French (2011).

2.1. Different goals of program evaluation—a broad brush comparison

It will be useful to reiterate a distinction that has been made elsewhere (see for example, Todd and Wolpin (2006) and Wolpin (2007)), between ex ante evaluation and ex post evaluation. *Ex post* policy evaluation occurs upon or after a policy has been implemented; information is collected about the outcomes experienced by those who participated in the "experiment" and an attempt is made to make inferences about the role of a treatment in influencing the outcomes. An ex post evaluation generally proceeds by selecting a statistical model with a tight fit to the experiment that actually happened (whether or not the experiment was "planned"). The claims that are licensed from such evaluations are context dependent—an experiment conducted among a specific

group of individuals, at a specific time and specific place, may or may not be a reliable indicator of what a treatment would do among a different group of individuals at a different time or place. The credibility of an ex post evaluation depends on the credibility of the statistical model of the *experiment*. Drug trials and social experiments are examples of "planned" experiments; similarly, regression discontinuity designs, although not necessarily planned, can also often provide opportunities for an ex post evaluation.

Ex ante evaluation, by contrast, does not require an experiment to have happened. It is the attempt to "study the effects of policy changes prior to their implementation" (Todd and Wolpin, 2006).[5] Unlike the ex post evaluation, the credibility of an ex ante evaluation depends on the credibility of the statistical model of the *behavior* of individuals and the *environment* to which the individuals are subjected. An influential ex ante evaluation was McFadden et al. (1977), which built a random utility model to forecast the demand for the San Francisco BART subway system before it was built. In that case, the random utility model is a more or less "complete", albeit highly stylized, description of utility maximizing agents, their "preferences", etc. In short, the statistical model explains *why* individuals make their observed choices. The model of behavior and the environment *is* the data generation process.

This contrasts sharply with ex post evaluation, where apart from the description of the treatment assignment mechanism, one is as agnostic as possible about what specific behavioral model is responsible for the observed data other than the assignment mechanism. We describe this below as "pan-theoretic"—the goal in an ex post evaluation is to write down a statistical model of the assignment process or the experiment that is consistent with as broad a class of potential models as possible. When the analyst has detailed institutional knowledge of the assignment mechanism, there is usually very little discretion in the choice of statistical model—it is dictated by the institutional details of the actual experiment. As observed by Wolpin (2007), however, this is not the case in the ex ante evaluation: "Researchers, beginning with the same question and using the same data, will generally differ along many dimensions in the modeling assumptions they make, and resulting models will tend to be indistinguishable in terms of model fit."

Since human behavior is so complicated and poorly understood (relative to the properties of simple treatment assignment mechanisms), ex ante evaluations typically place a high premium on some form of "parsimony"—some potential empirical pathways are necessarily omitted from the model. Researchers in different fields, or different economists, may construct models of the same outcomes which are very

[5] "Structural models" more generally refer to a collection of stylized mathematical descriptions of behavior and the environment which are combined to produce predictions about the effects of different choices, etc. It is a very broad area, and we make no attempt to review this literature. For a tiny sample of some of the methodological discussion, see Haavelmo (1944), Marschak (1953), Lucas (1976), Ashenfelter and Card (1982), Heckman (1991), Heckman (2000), Reiss and Wolak (2007), Heckman and Vytlacil (2007a), Deaton (2008), Fernández-Villaverde (2009), Heckman and Urzua (2009), and Keane (2009). We also ignore other types of structural models including "agent based" models (Windrum et al., 2007; Tesfatsion, 2007).

different. Because many different models—with different implications, but roughly the same "fit" to the data—might be used in an ex ante evaluation, there are a wide variety of ways in which such models are validated (see Heckman (2000), Keane and Wolpin (2007), Keane (2009) and the references therein for useful discussion). Given the goal of providing a good model of what might happen in contexts *different* than those in which the data was collected, testing or validating the model is considerably more difficult. Indeed, "the examination of models' predictive ability is not especially common in the microeconometrics literature" (Fang et al., 2007). Part of the difficulty is that by necessity, some variables in the model are "exogenous" (determined outside the model), and if these variables affect the outcome being studied, it is not sufficient to know the structure. For the ex ante evaluation to be reliable, "it is also necessary to know past and future values of all exogenous variables" (Marschak, 1953). Finally, it is worth noting that an ex ante *evaluation* (as opposed to a mere forecasting exercise) generally requires a specification of "values" (a clear discussion of the many issues involved can be found in Heckman and Smith (1998)).

In the following table, we outline some of the similarities and differences between the two kinds of evaluations, acknowledging the difficulties of "painting with a broad brush":

Ex post program evaluation	**Ex ante program evaluation**
What did the program do? Retrospective: what happened?	What do we think a program will do? Prospective/predictive: what would happen?
Focus on the program at hand	Focus on forecasting effects of different program
For what population *do* we identify causal effect?	For what population do we *want* to identify causal effect?
Desirable to have causal inferences not reliant on specific structural framework/model	Question ill-posed without structural framework/paradigm
No value judgments on "importance" of causal facts	Some facts will be more helpful than others
Inferences require assumptions	Predictions require assumptions
Desirable to test assumptions whenever possible	Desirable to test assumptions whenever possible
Ex Ante problem guides what programs to design/analyze	Would like predictions consistent with results of Ex Post evaluation
Inference most appropriate for situations that "resemble" the experiment and are similar to that which produce the observed data	Inferences intended for situations that are different than that which produced the observed data

2.2. The challenges of the ex post (descriptive) evaluation problem

Here we describe a prototypical ex post program evaluation, where the perspective is that an event has occurred (i.e. some individuals were exposed to the program, while others were not) and data has been collected. The ex post evaluation question is: Given the particular program that was implemented, and the data that was collected, what is the causal effect of the program on a specific outcome of interest?

For example, suppose a state agency implements a new program that requires unemployment insurance claimants to be contacted via telephone by a job search counselor for information and advice about re-employment opportunities, and data is collected on the labor market behavior of the claimants before and after being exposed to this program. The ex post evaluation problem is to assess the impact of this particular job search program on labor market outcomes (e.g. unemployment durations) for the population of individuals to whom it was exposed.

One might also want to know what the program's impact *would be* in a different state, or 5 years from now, or for a different population (e.g. recent high school graduates, rather than the recently unemployed), or if the job counselor were to make a personal visit to the UI claimant (rather than a phone call). But in our hypothetical example none of these things happened. We consider these questions to be the concern of an *ex ante* program evaluation—a forecast of the effect of a program that *has not* occurred. For now, we consider the program that was *actually implemented*, and its effect on the population to which the program was *actually* exposed, and focus on the goal of making as credible and precise causal inferences as possible (see Heckman and Vytlacil (2007a,b), Abbring and Heckman (2007), Keane and Wolpin (2007) and Todd and Wolpin (2006) for discussion).

We describe the general evaluation problem using the following notation:

- Y is the outcome of interest.
- D is the program, or treatment, status variable, equal to 1 if "treated" and 0 if not.
- W is a vector of all variables that could impact Y—some observable and others unobservable to the researcher—realized *prior* to the determination of program status. For example, W can represent immutable characteristics (e.g. race), constraints faced by, actions taken by, or information known to the individual. For any given individual, D and W, together, determine Y.
- U is a fundamentally unobservable random variable that indexes an individual's "type". By "type" we mean those individuals with both (1) exactly the same structural relation between D, W and Y, and (2) the exactly same value of W. It is important to note that this still allows (1) individuals with different Us to have the same value of W (as long as they have a different structural relation between D, W and Y), or (2) individuals with different Us to have the same relation between D, W and Y (as long as they have different values of W). Furthermore, this indexing says nothing about the

variability of variables outside of W (e.g. other factors that could exclusively influence treatment assignment) which could exist for those with the same U. $F_U(u)$ is the cdf of U.

A general framework for the evaluation problem can be given by the system:

$$W \equiv w(U) \tag{1}$$
$$P^* \equiv p^*(W, U) \equiv \Pr[D = 1|W, U] \tag{2}$$
$$Y \equiv y(D, W, U). \tag{3}$$

In the first equation, W is a random vector because U denotes the type of a randomly chosen individual from the population. With $w(\cdot)$ being a real-valued function, those with the same U (identical agents) will have the same W, but there may be variation in U conditional on an observed value of W : $w(\cdot)$ need not be one-to-one. Furthermore, since W is determined before D, D does not enter the function $w(\cdot)$.

The second equation defines the *latent* propensity to be treated, P^*. Program status can be influenced by type U or the factors W. Additionally, by allowing P^* to take values between 0 and 1, we are allowing for the possibility of "other factors" outside of W and U that could have impacted program status. If there are no "other factors", then P^* takes on the values 0 or 1. Even though our definition of types implies no variation in W conditional on U, it is still meaningful to consider the structural relation between W and P^*. In particular, for a given value of U equal to u, if one could select all alternative values of U for which the relation between W and P^* is exactly the same, then for that subset of Us, the variation in W could be used to trace out the impact of W on P^* for $U = u$. W might include years of education obtained prior to exposure to the job search assistance program, and one could believe that education could impact the propensity to be a program participant. It is important to note that P^* is quite distinct from the well-known "propensity score", as we will discuss in Section 4.3. Not only is P^* potentially a function of some unobservable elements of W, but even conditional on W, P^* can vary across individuals.

The final equation is the outcome equation, with the interest centering on the impact of D on Y, keeping all other things constant. As with P^*, although our definition of types implies no variation in W conditional on U, it is still meaningful to consider the structural relation between W and Y. Specifically, given a particular $U = u$, if one could select all the alternate values of U such that the relation between D, W and Y is the same, then for that subset of Us, the variation in W could be used to trace out the impact of W on Y for $U = u$.

Note that this notation has a direct correspondence to the familiar "potential outcomes framework" (Splawa-Neyman et al., 1990; Rubin, 1974; Holland, 1986).[6]

[6] Y_1 and Y_0 (in the potential outcomes framework) correspond to $y(1, w(U), U)$ and $y(0, w(U), U)$.

The framework also accommodates standard latent variable threshold-crossing models (Heckman, 1974, 1976, 1978) such as:

$$Y = \alpha + D\beta + X\gamma + \epsilon$$
$$D = 1[X\delta + V > 0]$$

where X, ϵ, (with an arbitrary joint distribution) are elements of W, and $P^* = \Pr[V > -X\delta | X, \epsilon]$. The framework also corresponds to that presented in Heckman and Vytlacil (2005).[7] The key difference is that we will not presume the existence of a continuously distributed instrument Z that is independent of all the unobservables in W.

Throughout this chapter, we maintain a standard assumption in the evaluation literature (and in much of micro-econometrics) that each individual's behaviors or outcomes do not directly impact the behaviors of others (i.e., we abstract from "peer effects", general equilibrium concerns, etc.).

Define the causal effect for an individual with $U = u$ and $W = w$ as

$$\Delta(w, u) \equiv y(1, w, u) - y(0, w, u).$$

If U and all the elements of W were observed, then the causal effect could be identified at any value of W and U provided there existed some treated and non-treated individuals.

The main challenge, of course, is that the econometrician will never observe U (even if individuals can be partially distinguished through the observable elements of W). Thus, even conditional on $W = w$, it is in general *only* possible to learn something about the *distribution* of $\Delta(w, U)$. Throughout this chapter we will focus on—as does much of the evaluation literature—*average* effects

$$\int \Delta(w(u), u)\psi(u)\mathrm{d}F_U(u) \tag{4}$$

where $\psi(u)$ is some weighting function such that $\int \psi(u)\mathrm{d}F_U(u) = 1$ (see Heckman and Vytlacil (2007a) and Abbring and Heckman (2007) for a discussion of distributional effects and effects other than the average).

The source of the causal inference problem stems from unobserved heterogeneity in P^*, which will cause treated and untreated populations to be noncomparable. The treated will tend to have higher P^* (and hence the U and W that lead to high P^*), while the untreated will have lower P^* (and hence values of U and W that lead to low P^*). Since U and W determine Y, the average Y will generally be different for different populations.

[7] Specifically, where we consider their X, U_1, and U_0 as elements of our vector W.

More formally, we have

$$\begin{aligned}
&E[Y|D=1]-E[Y|D=0]\\
&\quad=\int E[y(1,w(U),U)|D=1,P^*=p^*]f_{P^*|D=1}(p^*)\mathrm{d}p^*\\
&\qquad-\int E[y(0,w(U),U)|D=0,P^*=p^*]f_{P^*|D=0}(p^*)\mathrm{d}p^*\\
&\quad=\int E[y(1,w(U),U)|P^*=p^*]f_{P^*|D=1}(p^*)\mathrm{d}p^*\\
&\qquad-\int E[y(0,w(U),U)|P^*=p^*]f_{P^*|D=0}(p^*)\mathrm{d}p^* \qquad (5)
\end{aligned}$$

where the $f_{P^*|D=d}(p^*)$ is the density of P^* conditional on $D = d$, and the second equality follows from the fact that $E[y(d, w(U), U)|D = d, P^* = p^*] = E[y(d, w(U), U)|P^* = p^*]$: for all observations with an identical probability of receiving treatment, the distribution of unobservables will be identical between $D = 1$ and $D = 0$ populations.[8] Importantly, any nontrivial marginal density $f_{P^*}(p^*)$ will necessarily lead to $f_{P^*|D=1}(p^*) \neq f_{P^*|D=0}(p^*)$.[9]

In our discussion below, we will point out how various research designs grapple with the problem of unobserved heterogeneity in P^*. In summary, in an ex post evaluation problem, the task is to translate whatever knowledge we have about the assignment mechanism into restrictions on the functions given in Eqs (1), (2), or (3), and to investigate, as a result, what causal effects can be identified from the data.

2.2.1. Criteria for internal validity and the role of economic theory

We argue that in an ex post evaluation of a program, the goal is to make causal inferences with a high degree of "internal validity": the aim is to make credible inferences and qualify them as precisely as possible. In such a descriptive exercise, the degree of "external validity" is irrelevant. On the other hand, "external validity" will be of paramount importance when one wants to make predictive statements about the impact of the same program on a different population, or when one wants to use the inferences to make guesses about the possible effects of a slightly different program. That is, we view "external validity" to be the central issue in an attempt to use the results of an ex post evaluation for an ex ante program evaluation; we further discuss this in the next section.

What constitutes an inference with high "internal validity"?[10] Throughout this chapter we will consider three criteria. The first is the extent to which there is a tight

[8] Formally, $F_{U|D=1,P^*=p^*}(u) = \frac{\Pr[D=1|U\leq u,P^*=p^*]F_{U|P^*=p^*}(u)}{\Pr[D=1|P^*=p^*]} = F_{U|P^*=p^*}(u)$, and similarly, $F_{U|D=0,P^*=p^*}(u) = F_{U|P^*=p^*}(u)$.

[9] From Bayes' rule we have $f_{P^*|D=1}(p^*) = \frac{\Pr[D=1|P^*=p^*]f_{P^*}(p^*)}{\Pr[D=1]} = \frac{p^* f_{P^*}(p^*)}{\Pr[D=1]}$, and $f_{P^*|D=0}(p^*) = \frac{(1-p^*)f_{P^*}(p^*)}{1-\Pr[D=1]}$.

[10] Campbell and Cook (1979) contains a discussion of various "threats" to internal validity.

correspondence between what we know about the assignment-to-treatment mechanism and our statistical model of the process. In some cases, the assignment mechanism might leave very little room as to how it is to be formally translated into a statistical assumption. In other cases, little might be known about the process leading to treatment status, leaving much more discretion in the hands of the analyst to model the process. We view this discretion as potentially expanding the set of "plausible" (yet different) inferences that can be made, and hence generating doubt as to which one is correct.

The second criterion is the broadness of the class of models with which the causal inferences are consistent. Ideally, one would like to make a causal inference that is consistent with any conceivable behavioral model. By this criterion, it would be undesirable to make a causal inference that is only valid if a very specific behavioral model is true, and it is unknown how the inferences would change under plausible deviations from the model in question.

The last criterion we will consider is the extent to which the research design is testable; that is, the extent to which we can treat the proposed treatment assignment mechanism as a null hypothesis that could, in principle, be falsified with data (e.g. probabilistically, via a formal statistical test).

Overall, if one were to adopt these three criteria, then a research design would have low "internal validity" when (1) the statistical model is not based on what is actually known about the treatment assignment mechanism, but based entirely on speculation, (2) inferences are known only to be valid for one specific behavioral model amongst many other plausible alternatives and (3) there is no way to test the key assumption that achieves identification.

What is the role of economic (for that matter, any other) theory in the ex post evaluation problem? First of all, economic theories motivate what outcomes we wish to examine, and what causal relationships we wish to explore. For example, our models of job search (see McCall and McCall (2008) for example) may motivate us to examine the impact of a change in benefit levels on unemployment duration. Or if we were interested in the likely impacts of the "program" of a hike in the minimum wage, economists are likely to be most interested in the impact on employment, either for the purposes of measuring demand elasticities, or perhaps assessing the empirical relevance of a perfectly competitive labor market against that of a market in which firms face upward-sloping labor supply curves (Card and Krueger, 1995; Manning, 2003).

Second, when our institutional knowledge does not put enough structure on the problem to identify any causal effects, then assumptions about individuals' behavior *must* be made to make any causal statement, however conditional and qualified. In this way, structural assumptions motivated by economic theory can help "fill in the gaps" in the knowledge of the treatment assignment process.

Overall, in an ex post evaluation, the imposition of structural assumptions motivated by economic theory is done out of necessity. The ideal is to conjecture *as little as*

possible about individuals' behavior so as to make the causal inferences valid under the broadest class of all possible models. For example, one could imagine beginning with a simple Rosen-type model of schooling with wealth maximization (Rosen, 1987) as a basis for empirically estimating the impact of a college subsidy program on educational attainment and lifetime earnings. The problem with such an approach is that this would raise the question as to whether the causal inferences entirely depend on that particular Rosen-type model. What if one added consumption decisions to the model? What about saving and borrowing? What if there are credit constraints? What if there are unpredictable shocks to non-labor income? What if agents maximize present discounted utility rather than discounted lifetime wealth? The possible permutations go on and on.

It is tempting to reason that we have no choice but to adopt a specific model of economic behavior and to admit that causal inferences are conditional only on the model being true; that the only alternative is to make causal inferences that depend on assumptions that we do not even know we are making.[11] But this reasoning equates the *specificity* of a model with its *completeness*, which we believe to be very different notions.

Suppose, for example—in the context of evaluating the impact of our hypothetical job search assistance program—that the type of a randomly drawn individual from the population is given by the random variable U (with a cdf $F_U(u)$), that W represents all the constraints and actions the individual takes prior to, and in anticipation of, the determination of participating in the program D, and that outcomes are determined by the system given by Eqs (1)–(3). While there is no discussion of utility functions, production functions, information sets, or discount rates, the fact is that this is a *complete* model of the data generating process; that is, we have enough information to derive expressions for the joint distribution of the observables (Y, D, W) from the primitives of $F_U(u)$ and (1)–(3). At the same time it is not a very *specific* (or economic) model, but in fact, quite the opposite: it is perhaps the most general formulation that one could consider. It is difficult to imagine any economic model—including a standard job search model—being inconsistent with this framework.

Another example of this can be seen in the context of the impact of a job training program on earnings. One of the many different economic structures consistent with (1)–(3) is a Roy-type model of self-selection (Roy, 1951; Heckman and Honore, 1990) into training.[12] The Roy-type model is certainly *specific*, assuming perfect foresight on earnings in both the "training" or "no-training" regimes, as well as income maximization behavior. If one obtains causal inferences in the Roy model framework, an open question would be how the inferences change under different theoretical frameworks (e.g. a job search-type model, where training shifts the wage offer distribution upward). But if we

[11] See Keane (2009); Rosenzweig and Wolpin (2000) for a discussion along these lines.

[12] W could be observable components of human capital, $p^*(w, u) = 1[y(1, w(u), u) - y(0, w(u), u) \geq 0]$.

can show that the causal inferences are valid within the more general—but nonetheless *complete*—formulation of (1)–(3), then we know the inferences will still hold under both the Roy-type model, a job search model, or any number of plausible alternative economic theories.

2.3. The "parameter of interest" in an ex ante (predictive) evaluation problem

We now consider a particular kind of predictive, or ex ante, evaluation problem: suppose the researcher is interested in predicting the effects of a program "out of sample". For example, the impact of the Job Corps Training program on the earnings of youth in 1983 in the 10 largest metropolitan areas in the US may be the focus of an ex post evaluation, simply because the data at hand comes from such a setting. But it is natural to ask any one or a combination of the following questions: What would be the impact today (or some date in the future)? What would be the impact of an expanded version of the program in more cities (as opposed to the limited number of sites in the data)? What would be the impact on an older group of participants (as opposed to only the youth)? What would be the impact of a program that expanded eligibility for the program? These are examples of the questions that are in the domain of an ex ante evaluation problem.

Note that while the ex post evaluation problem has a descriptive motivation—the above questions implicitly have a prescriptive motivation. After all, there seems no other practical reason why knowing the impact of the program "today" would be any "better" than knowing the impact of the program 20 years ago, other than because such knowledge helps us make a particular policy decision today. Similarly, the only reason we would deem it "better" to know the impact for an older group of participants, or participants from less disadvantaged backgrounds, or participants in a broader group of cities is because we would like to evaluate whether actually targeting the program along any of these dimensions would be a good idea.

One can characterize an important distinction between the ex post and ex ante evaluation problems in terms of Eq. (4). In an ex post evaluation, the weights $\psi(u)$ are dictated by the constraints of the available data, and what causal effects are most plausibly identified. It is simply accepted as a fact—however disappointing it may be to the researcher—that there are only a few different weighted average effects that can be plausibly identified, whatever weights $\psi(u)$ they involve. By contrast, in an ex ante evaluation, the weights $\psi(w)$ are *chosen* by the researcher, irrespective of the feasibility of attaining the implied weighted average "of interest". These weights may reflect the researcher's subjective judgement about what is an "interesting" population to study. Alternatively, they may be implied by a specific normative framework. A clear example of the latter is found in Heckman and Vytlacil (2005), who begin with a Benthamite social welfare function to define a "policy relevant treatment effect", which is a weighted average treatment effect with a particular form for the weights $\psi(u)$.

One can thus view "external validity" to be the degree of similarity between the weights characterized in the ex post evaluation and the weights defined as being "of interest" in an ex ante evaluation. From this perspective, any claim about whether a particular causal inference is "externally valid" is necessarily imprecise without a clear definition of the desired weights and their theoretical justification. Again, the PRTE of Heckman and Vytlacil (2005) is a nice example where such a precise justification is given.

Overall, in contrast to the ex post evaluation, the goals of an ex ante evaluation are not necessarily tied to the specific context of or data collected on any particular program. In some cases, the researcher may be interested in the likely effects of a program on a population for which the program was already implemented; the goals of the ex post and ex ante evaluation would then be similar. But in other cases, the researcher may have reason to be interested in the likely effects of the program on different populations or in different "economic environments"; in these cases ex post and ex ante evaluations—even when they use the same data—would be expected to yield different results. It should be clear that however credible or reliable the ex post causal inferences are, ex ante evaluations using the same data will necessarily be more speculative and dependent on more assumptions, just as forecasting out of sample is a more speculative exercise than within-sample prediction.

2.3.1. Using ex post evaluations for ex ante predictions

In this chapter, we focus most of our attention on the goals of the ex post evaluation problem, that of achieving a high degree of internal validity. We recognize that the weighted average effects that are often identified in ex post evaluation research designs may not correspond to a potentially more intuitive "parameter of interest", raising the issue of "external validity". Accordingly—using well-known results in the econometric and evaluation literature—we sketch out a few approaches for extrapolating from the average effects obtained from the ex post analysis to effects that might be the focus of an ex ante evaluation.

Throughout the chapter, we limit ourselves to contexts in which a potential instrument is binary, because the real-world examples where potential instruments have been explicitly or "naturally" randomized, the instrument is invariably binary. As is well-understood in the evaluation literature, this creates a gap between what causal effects we *can* estimate and the potentially more "general" average effects of interest. It is intuitive that such a gap would diminish if one had access to an instrumental variable that is continuously distributed. Indeed, as Heckman and Vytlacil (2005) show, when the instrument Z is essentially randomized (and excluded from the outcome equation) *and* continuously distributed in such a way that $\Pr[D = 1|Z = 1]$ is continuously distributed on the unit interval, then the full set of what they define as Marginal Treatment Effects (MTE) can be used construct various policy parameters of interest.

3. RESEARCH DESIGNS DOMINATED BY KNOWLEDGE OF THE ASSIGNMENT PROCESS

In this section, we consider a group of research designs in which the model for the data generating process is to a large extent dictated by explicit institutional knowledge of how treatment status was assigned. We make the case that these four well-known cases deliver causal inferences with a high degree of "internal validity" because of at least three reasons: (1) some important or all aspects of the econometric model is a literal description of the treatment assignment process, (2) the validity of the causal inferences hold true within a seemingly broad class of competing behavioral models, and perhaps most importantly, (3) the statistical statements that describe the assignment process simultaneously generate strong observable predictions in the data. For these reasons, we argue that these cases might be considered "high-grade" experiments/natural experiments.[13]

In this section, we also consider the issue of "external validity" and the ex ante evaluation problem. It is well understood that in the four cases below, the populations for which average causal effects are identified may not correspond to the "populations of interest". The ATE identified in a small, randomized experiment does not necessarily reflect the impact of a widespread implementation of the program; the Local Average Treatment Effect (LATE) of Imbens and Angrist (1994) is distinct from the ATE; the causal effect identified by the Regression Discontinuity Design of Thistlethwaite and Campbell (1960) does not reflect the effect of making the program available to individuals whose assignment variable is well below the discontinuity threshold. For each case, we illustrate how imposing some structure on the problem can provide an explicit link between the quantities identified in the ex post evaluation and the parameters of interest in an ex ante evaluation problem.

3.1. Random assignment with perfect compliance

3.1.1. Simple random assignment

We start by considering simple random assignment with perfect compliance. "Perfect compliance" refers to the case that individuals who are assigned a particular treatment, do indeed receive the treatment. For example, consider a re-employment program for unemployment insurance claimants, where the "program" is being contacted (via telephone and/or personal visit) by a career counselor, who provides information that facilitates the job search process. Here, participation in this "program" is not voluntary, and it is easy to imagine a public agency randomly choosing a subset of the population of UI claimants to receive this treatment. The outcome might be time to re-employment or total earnings in a period following the treatment.

In terms of the framework defined by Eqs (1)–(3), this situation can be formally represented as

- D1: (Simple Random Assignment): $P^* = p_0,\ p_0 \in (0, 1)$, a nonrandom constant

[13] A discussion of "low grade" experiments can be found in Keane (2009). See also Rosenzweig and Wolpin (2000).

That is, for the entire population being studied, every individual has the same probability of being assigned to the program.

It is immediately clear that the distribution of P^* becomes degenerate, with a single mass point at $P^* = p_0$, and so the difference in the means in Eq. (5) becomes

$$\begin{aligned} E[Y|D=1] - E[Y|D=0] &= E[y(1, w(U), U)] - E[y(0, w(U), U)] \\ &= E[\Delta(w(U), U)] \\ &= \int \Delta(w(u), u) \mathrm{d}F_U(u) \equiv \text{ATE} \end{aligned}$$

where the ATE is the "average treatment effect". The weights from Eq. (4) are $\psi(u) = 1$ in this case. A key problem posed in Eq. (5) is the potential relationship between the latent propensity P^* and Y (the functions $E[y(1, w(U), U)|P^* = p^*]$ and $E[y(0, w(U), U)|P^* = p^*]$). Pure random assignment "solves" the problem by *eliminating* all variation in P^*.

Internal validity: pan-theoretic causal inference

Let us now assess this research design on the basis of the three criteria described in Section 2.2.1. First, given the general formulation of the problem in Eqs (1)–(3), Condition D1 is much less an assumption, but rather a literal description of the assignment process—the "D" denotes a *descriptive* element of the data generating process. Indeed, it is not clear how else one would formally describe the randomized experiment.

Second, the causal inference is apparently valid for any model that is consistent with the structure given in Eqs (1)–(3). As discussed in Section 2.2.1, it is difficult to conceive of a model of behavior that would *not* be consistent with (1), (2), and (3). So even though we are not explicitly laying out the elements of a specific model of behavior (e.g. a job search model), it should be clear that given the distribution $F_U(U)$, Eqs (1)–(3), and Condition D1 constitutes a *complete* model of the data generating process, and that causal inference is far from being "atheoretic". Indeed, the causal inference is best described as "pan-theoretic", consistent with a broad—arguably the broadest—class of possible behavioral models.

Finally, and perhaps most crucially, even though one could consider D1 to be a *descriptive* statement, we could alternatively treat it as a *hypothesis*, one with testable implications. Specifically, D1 implies

$$\begin{aligned} F_{U|D=1}(u) &= \frac{\Pr[D=1|U \leq u] F_U(u)}{\Pr[D=1]} \\ &= \frac{p_0 F_U(u)}{\Pr[D=1]} = F_U(u) \end{aligned} \tag{6}$$

and similarly, $F_{U|D=0}(u) = F_U(u)$. That is, the distribution of unobserved "types" is identical in the treatment and control groups. Since U is unobservable, this itself

is not testable. But a direct consequence of the result is that the pre-determined characteristics/actions must be identical between the two groups as well,

$$\begin{aligned} F_{W|D=d}(w) &= \Pr[w(U) \leq w | D = d] \\ &= \Pr[w(U) \leq w] \end{aligned}$$

which *is* a testable implication (as long as there are some observable elements of W).

The implication that the entire joint distribution of *all* pre-determined characteristics be identical in both the treatment and control states is indeed quite a stringent test, and also independent of any model of the determination of W. It is difficult to imagine a more stringent test.

Although it may be tempting to conclude that "even random assignment must assume that the unobservables are uncorrelated with treatment", on the contrary, the key point here is that the balance of unobservable types U between the treatment and control groups is not a primitive *assumption*; instead, it is a direct *consequence* of the assignment mechanism, which is described by D1. Furthermore, balance in the observable elements of W is *not* an additional assumption, but a natural implication of balance in the unobservable type U.

One might also find D1 "unappealing" since mathematically it seems like a strong condition. But from an ex post evaluation perspective, whether D1 is a "strong" or "weak" condition is not as important as the fact that D1 is beyond *realistic*: it is practically a literal description of the randomizing process.

3.1.2. Stratified/block randomization

Now, suppose there is a subset of elements in W—call this vector X—that are observed by the experimenter. A minor variant on the above mechanism is when the probability of assignment to treatment is different for different groups defined by X, but the probability of treatment is identical for all individuals *within* each group defined by X.[14] In our hypothetical job search assistance experiment, we could imagine initially stratifying the study population by their previous unemployment spell history: "short", "medium", and "long"-(predicted) spell UI claimants. This assignment procedure can be described as

- D2: (Random Assignment Conditional on X) $P^* = p^*(X),\ p^*(x) \in (0, 1)\ \forall x$

In this case, where there may be substantial variation in the unobservable type U for a given X, the probability of receiving treatment is identical for everyone with the same X.

[14] While this setup has been described as the "selection on observables", "potential outcomes", "switching regressions" or "Neyman-Rubin-Holland model" (Splawa-Neyman et al., 1990; Lehmann and Hodges, 1964; Quandt, 1958, 1972; Rubin, 1974; Barnow et al., 1976; Holland, 1986), to avoid confusion we will reserve the phrase "selection on observables" for the case where the investigator does not have detailed institutional knowledge of the selection process and treat the stratified/block randomization case as special case of simple randomization.

The results from simple random assignment naturally follow,

$$E[Y|D=1, X=x] - E[Y|D=0, X=x] = E[\Delta(W,U)|X=x] = \int \Delta(w(u),u)\mathrm{d}F_{U|X=x}(u),$$

essentially an average treatment effect, conditional on $X = x$.

We mention this case not because D2 is a weaker, and hence more palatable assumption, but rather, it is useful to know that the statement in D2—like the mechanism described by D1—is one that typically occurs when randomized experiments are implemented. For example, in the Negative Income Tax Experiments (Robins, 1985; Ashenfelter and Plant, 1990), X were the pre-experimental incomes, and families were randomized into the various treatment groups with varying probabilities, but those probabilities were identical for every unit with the same X. Another example is the Moving to Opportunity Experiment (Orr et al., 2003), which investigated the impact of individuals moving to a more economically advantaged neighborhood. The experiment was done in 5 different cities (Baltimore, Boston, Chicago, Los Angeles, and New York) over the period 1994-1998. Unanticipated variation in the rate at which people found eligible leases led them to change the fraction of individuals randomly assigned to the treatments two different times during the experiment (Orr et al., 2003, page 232). In this case, families were divided into different "blocks" or "strata" by location × time and there was a different randomization ratio for each of these blocks.

This design—being very similar to the simple randomization case—would have a similar level of internal validity, according to two of our three criteria. Whether this design is testable (the third criterion we are considering) depends on the available data. By the same argument as in the simple random assignment case, we have

$$F_{W|D=d,X=x}(w) = \Pr[w(U) \leq w|D=d, X=x] = \Pr[w(U) \leq w|X=x].$$

So if the conditional randomization scheme is based on *all* of the Xs that are observed by the analyst, then there are no testable implications. On the other hand, if there are additional elements in W that are observed (but not used in the stratification), then once again, one can treat D2 as a hypothesis, and test that hypothesis by examining whether the distribution of those extra variables are the same in the treated and control groups (conditional on X).

3.1.3. The randomized experiment: pre-specified research design and a chance setup

We have focused so far on the role that randomization (as described by D1 or D2) plays in ensuring a balance of the unobservable types in the treated and control groups, and

have argued that in principle, this can deliver causal inferences with a high degree of internal validity.

Another characteristic of the randomized experiment is that it can be described as "pre-specified" research design. In principle, *before* the experiment is carried out, the researcher is able to dictate in advance what analyses are to be performed. Indeed, in medical research conducted in the US, prior to conducting an medical experiment, investigators will frequently post a complete description of the experiment in advance at a web site such as clincaltrials.gov. This posting includes how the randomization will be performed, the rules for selecting subjects, the outcomes that will be investigated, and what statistical tests will be performed. Among other things, such pre-announcement prevents the possibility of "selective reporting"—reporting the results only from those trials that achieve the "desired" result. The underlying notion motivating such procedure has been described as providing a "severe test"—a test which "provides an overwhelmingly good chance of revealing the presence of a specific error, if it exists—but not otherwise" (Mayo, 1996, page 7). This notion conveys the idea that convincing statistical evidence does not rely *only* on the "fit" of the data to a particular hypothesis but on the *procedure* used to arrive at the result. Good procedures are ones that make fewer "errors."

It should be recognized, of course, that this "ideal" of pre-specification is rarely implemented in social experiments in economics. In the empirical analysis of randomized evaluations, analysts often cannot help but be interested in the effects for different subgroups (in which they were not initially interested), and the analysis can soon resemble a data-mining exercise.[15] That said, the problem of data-mining is not specific to randomized experiments, and a researcher armed with a lot of explanatory variables in a non-experimental setting can easily find many "significant" results even among purely randomly generated "data" (see Freedman (1983) for one illustration). It is probably constructive to consider that there is a spectrum of pre-specification, with the pre-announcement procedure described above on one extreme, and specification searching and "significance hunting" with non-experimental data on the other. In our discussion below, we make the case that detailed knowledge of the assignment-to-treatment process can serve much the same role as a pre-specified research design in "planned" experiments—as a kind of "straight jacket" which largely dictates the nature of statistical analysis.

Another noteworthy consequence of this particular data generating process is that it is essentially a "statistical machine" or a "chance set up" (Hacking, 1965) whose "operating characteristics" or statistical properties are well-understood, such as a coin flip. Indeed, after a randomizer assigns n individuals to the (well-defined) treatment, and n individuals to the control, for a total of $N = 2n$ individuals, one can conduct a non-parametric *exact* test of a sharp null hypothesis that does not require *any* particular distributional assumptions.

15 See Deaton (2008).

Consider the sharp null hypothesis that there is no treatment effect for any individuals (which implies that the two samples are drawn from the same distribution). In this case the assignment of the label "treatment" or "control" is arbitrary. In this example there are $P = \binom{2n}{n}$ different ways the labels "treatment" and "control" *could have* been assigned. Now consider the following procedure:

1. Compute the difference in means (or any other interesting test statistic). Call this $\widehat{\Delta}$.
2. Permute the label treatment or control and compute the test statistic under this assignment of labels. This will generate P different values of the test statistic Δ_p^* for $p = 1, \ldots, P$. These collection of these observations yield an exact distribution of the test statistic.
3. One can compute the p-value such that the probability that a draw from this distribution would exceed $|\hat{\Delta}|$.

This particular "randomization" or "permutation" test was originally proposed by Fisher (1935) for its utility to "supply confirmation whenever, rightly or, more often wrongly, it is suspected that the simpler tests have been appreciably injured by departures from normality." (Fisher, 1966, page 48) (see Lehmann (1959, pages 183–192) for a detailed discussion). Our purpose in introducing it here is *not* to advocate for randomization inference as an "all purpose" solution for hypothesis testing; rather our purpose is to show just how powerful detailed institutional knowledge of the DGP can be.

3.1.4. Ex ante evaluation: predicting the effects of an expansion in the program

Up to this point, with our focus on an ex post evaluation we have considered the question, "For the individuals exposed to the randomized evaluation, what was the impact of the program?" We now consider a particular ex ante evaluation question, "What would be the impact of a full-scale implementation of the program?", in a context when that full-scale implementation has not occurred. It is not difficult to imagine that the individuals who participate in a small-scale randomized evaluation may differ from those who would receive treatment under full-scale implementation. One could take the perspective that this therefore makes the highly credible/internally valid causal inferences from the randomized evaluation irrelevant, and hence that there is no choice but to pursue non-experimental methods, such as a structural modeling approach to evaluation, to answer the "real" question of interest.

Here we present an alternative view that this ex ante evaluation question is an extrapolation problem. And far from being irrelevant, estimates from a randomized evaluation can *form the basis* for such an extrapolation. And rather than viewing structural modeling and estimation as an alternative or substitute for experimental methods, we consider the two approaches to be potentially quite complementary in carrying out this extrapolation. That is, one can adopt certain assumptions about behavior and the

structure of the economy to make precise the linkage between highly credible impact estimates from a small-scale experiment and the impact of a hypothetical full-scale implementation.

We illustrate this with the following example. Suppose one conducted a small-scale randomized evaluation of a job training program where participation in the *experimental study* was voluntary, while actual receipt of training was randomized. The question is, what would be the impact on earnings if we opened up the program so that participation in the *program* was voluntary?

First, let us define the parameter of interest as

$$E\left[Y^T\right] - E\left[Y^N\right]$$

where Y^T and Y^N are the earnings of a randomly drawn individual under two regimes: full-scale implementation of the program (T), or no program at all (N). This corresponds to the parameter of interest that motivates the Policy Relevant Treatment Effect (PRTE) of Heckman and Vytlacil (2001b). We might like to know the average earnings gain for everyone in the population. We can also express this as

$$\begin{aligned} E\left[Y^T\right] - E\left[Y^N\right] &= E\left[D^T Y_1^T + \left(1 - D^T\right) Y_0^T\right] - E\left[Y_0^N\right] \\ &= E\left[Y_1^T | D^T = 1\right] \Pr\left[D^T = 1\right] \\ &\quad + E\left[Y_0^T | D^T = 0\right] \Pr\left[D^T = 0\right] - E\left[Y_0^N\right] \end{aligned} \tag{7}$$

where the D^T is the treatment status indicator in the T regime, and the subscripts denote the potential outcomes.

Make the following assumptions:

- S1 (Linear Production Technology): $Q = \sum_{j=1}^{K} a(j)L(j)$, where Q is the amount of output, $L(j)$ is the total amount of labor supplied by workers with j units of human capital, and $a(j)$ are technological parameters with $a(j+1) > a(j)$.
- S2 (Job Training as Human Capital): the random variable J is the individual's endowment of human capital, and Δ is the gain in human capital due to training, so that in the implementation regime, human capital is $J + \Delta D^T$.
- S3 (Inelastic Labor Supply): each individual inelastically supplies L units of labor.
- S4 (Profit maximizing price-taking): $W = a(J + \Delta \cdot D^T)$.

This setup will imply that

$$Y_0^N = W \cdot L = a(J) \cdot L = Y_0^T.$$

Thus, S1 through S4 are simply a set of economic assumptions that says potential outcomes are unaffected by the implementation of the program; this corresponds to what Heckman and Vytlacil (2005) call *policy invariance*. This policy invariance comes about because of the linear production technology, which implies that wages are determined by the technological parameters, and *not* the supply of labor for each level of human capital.

With this invariance, we may suppress the superscript T for the potential outcomes; Eq. (7) will become

$$\begin{aligned}&\Big(E\Big[Y_1|D^T=1\Big]-E\Big[Y_0|D^T=1\Big]\Big)\Pr\Big[D^T=1\Big]\\&=\Big(E\Big[Y_1-Y_0|D^T=1\Big]\Big)\Pr\Big[D^T=1\Big].\end{aligned}$$

[16] Note that the key causal parameter $E[Y_1 - Y_0|D^T = 1]$ will in general be different from $E[Y_1 - Y_0|D^E = 1]$, where D^E is the indicator for having participated in the smaller scale randomized experiment (bearing the risk of not being selected for treatment). That is, the concern is that those who participate in the experimental study may not be representative of the population that would eventually participate in a full-scale implementation.

How could they be linked? Consider the additional assumptions

- S5 (Income Maximization; Perfect Information; Selection on Gains): $D^T = 1$ iff $Y_1 - Y_0 > c_f + c_p$, with c_f is the "fixed" cost to applying for the program, and c_p is the monetary cost to the individual from receiving the treatment.
- S6 (Risk Neutrality) Individuals maximize expected income.

Together, S5 and S6 imply that we could characterize the selection into the program in the experimental regime as

$$D^E = 1 \quad \text{iff} \quad p(Y_1 - Y_0) > c_f + pc_p$$

where p is the probability of being randomized into receiving the treatment (conditional on participating in the experimental study). Note that this presumes that in the experimental regime, all individuals in the population have the option of signing up for the experimental evaluation.

Finally, assume a functional form for the distribution of training effects in the population:

- S7 (Functional Form: Normality): $Y_1 - Y_0$ is normally distributed with variance σ^2.

[16] This is because $E[Y_0^N] = E[Y_0^N|D^T = 1]\Pr[D^T = 1] + E[Y_0^N|D^T = 0]\Pr[D^T = 0]$.

Applying assumption S7 yields the following expressions

$$E\Big[Y_1 - Y_0 | D^T = 1\Big] = E[Y_1 - Y_0] + \sigma \frac{\phi\left(\frac{c_f + c_p - E[Y_1 - Y_0]}{\sigma}\right)}{1 - \Phi\left(\frac{c_f + c_p - E[Y_1 - Y_0]}{\sigma}\right)} \tag{8}$$

$$E\Big[Y_1 - Y_0 | D^E = 1\Big] = E[Y_1 - Y_0] + \sigma \frac{\phi\left(\frac{c_p + \frac{c_f}{p} - E[Y_1 - Y_0]}{\sigma}\right)}{1 - \Phi\left(\frac{c_p + \frac{c_f}{p} - E[Y_1 - Y_0]}{\sigma}\right)}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal pdf and cdf, respectively.

The probability of assignment to treatment in the experimental regime, p, characterizes the scale of the program. The smaller p is, the smaller the expected gain to participating in the experimental study, and hence the average effect of the study participants will be more positively selected. On the other hand, as p approaches 1, the experimental estimate approaches the policy parameter of interest because the experiment *becomes* the program of interest. Although we are considering the problem of predicting a "scaling up" of the program, this is an interesting case to case to consider because it implies that for *an already existing* program, one can potentially conduct a randomized evaluation, where a small fraction of individuals are denied the program (p close to 1), and the resulting experimentally identified effect $E[Y_1 - Y_0 | D^E = 1]$ can be directly used to predict the aggregate impact of completely shutting down the program.[17]

The left-hand side of the first equation is the "parameter of interest" (i.e. what we want to know) in an ex ante evaluation problem. The left-hand side of the second equation is "what can be identified" (i.e. what we do know) from the experimental data in the ex post evaluation problem. The latter may not be "economically interesting" per se, but at the same time it is far from being unrelated to the former.

Indeed, the average treatment effect identified from the randomized experiment is the starting point or "leading term", when we combine the above two expressions to yield

$$E\Big[Y_1 - Y_0 | D^T = 1\Big] = E\Big[Y_1 - Y_0 | D^E = 1\Big] + \sigma \left[\frac{\phi\big(\Phi^{-1}\big(1 - \Pr[D^T = 1]\big)\big)}{\Pr[D^T = 1]} - \frac{\phi\big(\Phi^{-1}\big(1 - \Pr[D^E = 1]\big)\big)}{\Pr[D^E = 1]}\right]$$

with the only unknown parameters in this expression being $\Pr[D^T = 1]$, the predicted take-up in a full-scale implementation, and σ, the degree of heterogeneity of the

[17] Heckman and Vytlacil (2005) make this point clearly, noting that the treatment on the treated parameter is the key ingredient to predicting the impacts of shutting down the program.

potential training effects *in the entire population*. It is intuitive that *any* ex ante evaluation of the full-scale implementation that has not yet occurred will, at a minimum, need these two quantities.

In presenting this example, we do not mean to assert that the economic assumptions S1 through S7 are particularly realistic. Nor do we assert they are minimally sufficient to lead to an extrapolative expression. There are as many different ways to model the economy as there are economists (and probably more!). Instead, we are simply illustrating that an ex ante evaluation attempt can directly *use* the results of an ex post evaluation, and in this way the description of the data generating process in an ex post evaluation (D1 or D2) can be quite complementary to the structural economic assumptions (S1 through S7). D1 is the key assumption that helps you identify whether there is credible evidence—arguably the most credible that is possible—of a causal phenomenon, while S1 through S7 provides a precise framework to think about making educated guesses about the effects of a program that has yet to be implemented. Although $E[Y_1 - Y_0|D^E = 1]$ may not be of direct interest, obtaining credible estimates of this quantity would seem helpful for making a prediction about $E[Y_1 - Y_0|D^T = 1]$.

3.2. Random assignment: imperfect compliance

We now consider another data generating process that we know often occurs in reality—when there is randomization in the "intent to treat", but where participation in the program is potentially non-random and driven by self-selection. To return to our hypothetical job search assistance program, instead of mandating the treatment (personal visit/phone call from a career counselor), one could make participation in receiving such a call voluntary. Furthermore, one could take UI claimants and randomize them into two groups: one group receives information about the existence of this program, and the other does not receive the information. One can easily imagine that those who voluntarily sign up to be contacted by the job counselor might be systematically different from those who do not, and in ways related to the outcome. One can also imagine being interested in knowing the "overall effect" of "providing information about the program", but more often it is the case that we are interested in participation in the program *per se* (the treatment of "being contacted by the job counselor").

We discuss this widely known data generating process within the very general framework described by Eqs (1)–(3). We will introduce a more accommodating monotonicity condition than that employed in Imbens and Angrist (1994) and Angrist et al. (1996). When we do so, the familiar "Wald" estimand will give an interpretation of an average treatment effect that is not quite as "local" as implied by the "local average treatment effect" (LATE), which is described as "the average treatment effect for the [subpopulation of] individuals whose treatment status is influenced by changing an exogenous regressor that satisfies an exclusion restriction" Imbens and Angrist (1994).

We begin with describing random assignment of the "intent to treat" as

- D3 (Random Assignment of Binary Instrument): $\Pr[Z = 1|U = u] = p_{z1} \in (0, 1)$, a nonrandom constant. We can thus write $P^* = p^*(W, U) = p_{z1}p_1^*(w(U), U) + (1 - p_{z1})p_0^*(w(U), U)$ where $p_z^*(w(u), u) \equiv \Pr[D = 1|U = u, Z = z]$.

This is analogous to D1 (and D2), except that instead of randomizing the treatment, we are randomizing the instrumental variable. Like D1 and D2, it is appropriate to consider this a *description* of the process when we know that Z has been randomized.

Since we have introduced a new variable Z, we must specify how it relates to the other variables:

- S8 (Excludability): $Y = y(D, W, U)$, $W = w(U)$ (Z is not included as an argument in either function).

Although this is a re-statement of Eqs (1) and (3), given the existence of Z, this is a substantive and crucial assumption. It is the standard excludability condition: Z cannot have an impact on Y, either directly or indirectly through influencing the other factors W. It is typically *not* a literal descriptive statement in the way that D1 through D3 can sometimes be. It is a structural ("S") assumption on the same level as S1 through S7 and it may or may not be plausible depending on the context.

Finally, we have

- S9 (Probabilistic Monotonicity): $p_1^*(w(u), u) \geq p_0^*(w(u), u)$ for all u.

S9 is a generalization of the monotonicity condition used in Imbens and Angrist (1994) and Angrist et al. (1996). In those papers, $p_1^*(w(u), u)$ or $p_0^*(w(u), u)$ take on the values 1 or 0; that is, for a given individual type U, their treatment status is *deterministic* for a given value of the instrument Z. This would imply that P^* would have a distribution with three points of support: 0 (the latent propensity for "never-takers"), p_{z1} (the latent propensity for "compliers"), and 1 (the latent propensity for "always-takers").[18]

In the slightly more general framework presented here, for each type U, for a given value of the instrument Z, treatment status is allowed to be *probabilistic*: some *fraction* (potentially strictly between 0 and 1) of them will be treated. P^* can thus take on a continuum of values between 0 and 1. The probabilistic nature of the treatment assignment can be interpreted in at least two ways: (1) for a particular individual of type U, there are random shocks beyond the individual's control that introduce some uncertainty into the treatment receipt (e.g. there was a missed newspaper delivery, so the individual did not see an advertisement for the job counseling program), or (2) even for the same individual type U (and hence with the same potential outcomes), there is

[18] Without the monotonicity condition, the other point of support would be $(1 - p_{z1})$, the latent propensity of the "defiers".

heterogeneity in individuals in the factors that determine participation even conditional on U (e.g. heterogeneity in costs of participation).

S9 allows some violations of "deterministic" monotonicity at the individual level (the simultaneous presence of "compliers" and "defiers"), but requires that—conditional on the individual type U—the *probability* of treatment rises when Z moves from 0 to 1. In other words, S9 requires that—conditional on U—on average the "compliers" outnumber the "defiers". To use the notation in the literature, where D_0 and D_1 are the possible treatments when $Z = 0$ or 1, respectively, the monotonicity condition discussed in the literature is $\Pr[D_1 > D_0] = 1$. By contrast, S9 requires $\Pr[D_1 > D_0|U] - \Pr[D_1 < D_0|U] \geq 0$. Integrating over U, S9 thus implies that $\Pr[D_1 > D_0] - \Pr[D_1 < D_0] \geq 0$, but the converse is not true. Furthermore, while $\Pr[D_1 > D_0] = 1$ implies S9, the converse is not true.

It follows that

$$E[y(D, w(u), u)|Z = z, U = u] = y(0, w(u), u) + p_z^*(w(u), u)\Delta(w(u), u).$$

Averaging over the distribution of U conditional on Z yields

$$\begin{aligned} &E[y(D, w(U), U)|Z = z] \\ &\quad = \int (y(0, w(u), u) + p_z^*(w(u), u)\Delta(w(u), u))\mathrm{d}F_{U|Z=z}(u). \end{aligned}$$

Taking the difference between the $Z = 0$ and $Z = 1$ individuals, this yields the reduced-form

$$\begin{aligned} &E[Y|Z = 1] - E[Y|Z = 0] \\ &\quad = \int \Delta(w(u), u)\left[p_1^*(w(u), u) - p_0^*(w(u), u)\right]\mathrm{d}F_U(u) \end{aligned}$$

where D3 allows us to combine the two integrals. Note also that without S8, we would be unable to factor out the term $\Delta(w(u), u)$.

It is useful here to contrast the DGP given by D3 and S8 with the randomized experiment with perfect compliance, in how it confronts the problem posed by Eq. (5). With perfect compliance, the randomization made it so that P^* was the same constant for both treated and control individuals, so the two terms in Eq. (5) could be combined. With non-random selection into treatment, we must admit the possibility of variability in P^*. But instead of making the contrast between $D = 1$ and $D = 0$, it is made between $Z = 1$ versus $Z = 0$ individuals, who, by D3, have the same distribution of types ($F_{U|Z=1}(u) = F_{U|Z=0}(u)$). Thus, the randomized instrument allows us to compare two groups with the same *distribution* of latent propensities P^*: $F_{P^*|Z=1}(p^*) = F_{P^*|Z=0}(p^*)$.

Dividing the preceding equation by a normalizing factor, it follows that the Wald Estimand will identify

$$\frac{E[Y|Z=1]-E[Y|Z=0]}{E[D|Z=1]-E[D|Z=0]} = \int \Delta(w(u),u)\frac{p_1^*(w(u),u)-p_0^*(w(u),u)}{E[D|Z=1]-E[D|Z=0]}\mathrm{d}F_U(u). \tag{9}$$

Therefore, there is an alternative to the interpretation of the Wald estimand as the LATE.[19] It can be viewed as the *weighted* average treatment effect for the entire population where the weights are proportional to the increase in the probability of treatment caused by the instrument, $p_1^*(w(u),u) - p_0^*(w(u),u)$.[20] This weighted average interpretation requires the weights to be non-negative, which will be true if and only if the probabilistic monotonicity condition S9 holds. Note the connection with the conventional LATE interpretation: when treatment is a deterministic function of Z, then the monotonicity means only the compliers (i.e. $p_1^*(w(u),u) - p_0^*(w(u),u) = 1$) collectively receive 100 percent of the weight, while all other units receive 0 weight.

The general framework given by Eqs (1)–(3), and the weaker monotonicity condition S9 thus leads to a less "local" interpretation than LATE. For example, Angrist and Evans (1998) use a binary variable that indicates whether the first two children were of the same gender (*Same Sex*) as an instrument for whether the family ultimately has more than 2 children (*More than 2*). They find a first-stage coefficient of around 0.06. The conventional monotonicity assumption, which presumes that D is a deterministic function of Z, leads to the interpretation that we know that 6 percent of families are "compliers": those that are induced to having a third child because their first two children were of the same gender. This naturally leads to the conclusion that the average effect "only applies" to 6 percent of the population.

In light of Eq. (9), however, an alternative interpretation is that the Wald estimand yields a weighted average of 100 percent of the population, with individual weights proportional to the individual-specific impact of (*Same Sex*) on (*More than 2*). In fact, if (*Same Sex*) had the same 0.06 impact on the probability of having more than 2 children *for all families*, the Wald Estimand will yield the ATE. Nothing in this scenario prevents substantial amount of variation in $p_1^*(w(U),U)$, $p_0^*(w(U),U)$, (and hence P^*), as well as non-random selection into treatment (e.g. correlation between P^* and $y(d,W,U)$).[21] With our hypothetical instrument of "providing information about the job counseling program", a first-stage effect on participation of 0.02 can be interpreted as a 0.02 effect in probability for *all individuals.*

[19] Note that $\int p_1^*(w(u),u) - p_0^*(w(u),u)\mathrm{d}F_U(u) = E[D|Z=1] - E[D|Z=0]$.

[20] Alternatively, one can view the weights as being proportional to the fraction of compliers in excess of the defiers among individuals of the same type U: $p_1^*(w(u),u) - p_0^*(w(u),u) = \Pr[D_1 > D_0|U=u] - \Pr[D_1 < D_0|U=u]$.

[21] But in this example, $p_0^*(w(U),U)$ would have to be bounded above by $1 - 0.06 = 0.94$.

In summary, the data generating process given by D3, S8, and S9 – compared to one where there is deterministic monotonicity—is a broader characterization of the models for which the Wald estimand identifies an average effect. Accordingly, the Wald estimand can have a broader interpretation as a weighted average treatment effect. The framework used to yield the LATE interpretation restricts the heterogeneity in P^* to have only three points of support, 0, p_{z1}, and 1. Thus, the LATE interpretation—which admits that effects can only be identified for those with $P^* = p_{z1}$ is one that most exaggerates the "local" or "unrepresentativesness" of the Wald-identified average effect.

Finally, it is natural to wonder why there is so much of a focus on the Wald estimand. In a purely ex post evaluation analysis, the reason is *not* that IV is a "favorite" or "common" estimand.[22] Rather, in an ex post evaluation, we may have limited options, based on the realities of how the program was conducted, and what data are available. So, for example, as analysts we may be confronted with an instrument, "provision of information about the job counseling program" (Z), which was indeed randomized as described by D3, and on purely theoretical grounds, we are comfortable with the additional structural assumptions S8 and S9. But suppose we are limited by the observable data (Y, D, Z), and know nothing else about the structure given in Eqs (1)–(3), and therefore wish our inferences to be invariant to any possible behavioral model consistent with those equations. If we want to identify some kind of average $\Delta(w(U), U)$, then what alternative do we have but the Wald estimand? It is not clear there is one.

The definition of the weights of "interest" is precisely the first step of an *ex ante* evaluation of a program. We argue that the results of an analysis that yields us an average effect, as in (9), may well not be the direct "parameter of interest", but could be used as an ingredient to predict such a parameter in an *ex ante* evaluation analysis. We illustrate this notion with a simple example below.

3.2.1. Assessment

In terms of our three criteria to assess internal validity, how does this research design fare—particularly in comparison to the randomized experiment with perfect compliance? First, only *part* of the data generating process given by D3, S8, and S9 is a literal description of the assignment process: if Z is truly randomly assigned, then D3 is not so much an assumption, but a description. On the other hand, S8 and S9 will typically be conjectures about behavior rather than being an implication of our institutional knowledge.[23] This is an example where there are "gaps" in our understanding of

[22] Heckman and Vytlacil (2005) and Heckman et al. (2006) correctly observe that from the perspective of the ex ante evaluation problem a singular focus on estimators without an articulated model will not, in general, be helpful in answering a question of economic interest. In an ex post evaluation, however, careful qualification of what parameters are identified from the experiment (as opposed to the parameters of more economic interest) is a desirable feature of the evaluation.

[23] The exception to this is that, in some cases, our institutional knowledge may lead us to know that those assigned $Z = 0$ are barred from receiving treatment. S9 will necessarily follow.

the assignment process, and structural assumptions work together with experimental variation to achieve identification.

As for our second criterion, with the addition of the structure imposed by S8 and S9, it is clear that the class of all behavioral models for which the causal inference in Eq. (9) is valid is smaller. It is helpful to consider, for our hypothetical instrument, the kinds of economic models that would or would not be consistent with S8 and S9. If individuals are choosing the best job search activity amongst all known available feasible options, then the instrument of "providing information about the existence of a career counseling program" could be viewed as adding one more known alternative. A standard revealed preference argument would dictate that if an individual already chose to participate under $Z = 0$, then it would still be optimal if $Z = 1$: this would satisfy S9. Furthermore, it is arguably true that most attempts at modeling this process would not specify a direct impact of this added information on human capital; this would be an argument for S8. On the other hand, what if the information received about the program carried a signal of some other factor? It could indicate to the individual, that the state agency is monitoring their job search behavior more closely. This might induce the individual to search more intensively, independently of participating in the career counseling program; this would be violation of the exclusion restriction S8. Or perhaps the information provided sends a positive signal about the state of the job market and induces the individual to pursue other job search activities instead of the program; this might lead to a violation of S9.

For our third criterion, we can see that some aspects of D3, S8, and S9 are potentially testable. Suppose the elements of W can be categorized into the vector W^- (the variables determined *prior* to Z) and W^+ (after Z). And suppose we can observe a subset of elements from each of these vectors as variables X^- and X^+, respectively. Then the randomization in D3 has the direct implication that the distributions of X^- for $Z = 1$ and $Z = 0$ should be identical:

$$F_{X^-|Z=1}(x) = F_{X^-|Z=0}(x).$$

Furthermore, since the exclusion restriction S8 dictates that all factors W that determine Y are not influenced by Z, then D3 and S8 jointly imply that the distribution of X^+ are identical for the two groups:

$$F_{X^+|Z=1}(x) = F_{X^+|Z=0}(x).$$

The practical limitation here is that this test pre-supposes the researcher's X^+ really do reflect elements of W^+ that influence Y. If X^+ are not a subset of W^+, then even if there is imbalance in X^+, S8 could still hold. Contrast this with the implication of D3 (and D1 and D2) that *any* variable determined prior to the random assignment should have a distribution that is identical between the two randomly assigned groups. Also, there seems no obvious way to test the proposition that Z does not directly impact Y, which is another condition required by S8.

Finally, if S9 holds, it must also be true that

$$\Pr\left[D = 1 | X^{-} = x, Z = 1\right] - \Pr\left[D = 1 | X^{-} = x, Z = 0\right] \geq 0, \quad \forall x.$$

That is, if probabilistic monotonicity holds for all U, then it must also hold for groups of individuals, defined by the value of X^{-} (which is a function of U). This inequality also holds for *any* variables determined prior to Z.

In summary, we conclude (unsurprisingly) that for programs where there is random assignment in the "encouragement" of individuals to participate, causal inferences will be of strictly lower internal validity, relative to the perfect compliance case. Nevertheless, the design does seem to satisfy—even if to a lesser degree—the three criteria that we are considering. Our knowledge of the assignment process does dictate an important aspect of the statistical model (and other aspects need to be modeled with structural assumptions), the causal inferences using the Wald estimand appear valid within a reasonably broad class of models (even it is not as broad as that for the perfect compliance case), and there are certain aspects of the design that generate testable implications.

3.2.2. Ex ante evaluation: extrapolating from LATE to ATE

Perhaps the most common criticism leveled at the LATE parameter is that it may not be the "parameter of interest".[24] By the same token, one may have little reason to be satisfied with the particular weights in the average effect expressed in (9). Returning to our hypothetical example in which the instrument "provide information on career counseling program" is randomized, the researcher may not be interested in an average effect that over-samples those who are more influenced by the instrument. For example, a researcher might be interested in predicting the average impact of individuals of a *mandatory* job counseling program, like the hypothetical example in the case of the randomized experiment with perfect compliance. That is, it may be of interest to predict what would happen if people were required to participate (i.e. every UI claimant will receive a call/visit from a job counselor). Moreover, it has been suggested that LATE is an "instrument-specific" parameter and a singular focus on LATE risks conflating "definition of parameters with issues of identification" (Heckman and Vytlacil, 2005): different instruments can be expected to yield different "LATEs".

Our view is that these are valid criticisms from the perspective of an ex ante evaluation standpoint, where causal "parameters of interest" are defined by a theoretical framework describing the policy problem. But from an ex post evaluation perspective, within which internal validity is the primary goal, these issues are, by definition, unimportant. When the goal is to describe whatever one can about the causal effects of a program that was actually implemented (e.g. randomization of "information about the job counseling

[24] Discussions on this point can be found, for example, in Heckman and Vytlacil (2001b), Heckman (2001), and Heckman and Vytlacil (2005), as well as Heckman and Vytlacil (2007a,b) and Abbring and Heckman (2007).

program", Z), a rigorous analysis will lead to precise statements about the causal phenomena that are *possible* to credibly identify. Sometimes what one *can* credibly identify may correspond to a desired parameter from a well-defined ex ante evaluation; sometimes it will not.

Although there has been considerable emphasis in the applied literature on the fact that LATE may differ from ATE, as well as discussion in the theoretical literature about the "merits" of LATE as a parameter, far less effort has been spent in actually using estimates of LATE to learn about ATE. Even though standard "textbook" selection models can lead to simple ways to extrapolate from LATE to ATE (see Heckman and Vytlacil (2001a) and Heckman et al. (2001, 2003)), our survey of the applied literature revealed very few other attempts to actually produce these extrapolations.

Although we have argued that the ATE from a randomized experiment may not directly correspond to the parameter of interest, for the following derivations, let us stipulate that ATE is useful, either for extrapolation (as illustrated in Section 3.1.4), or as an "instrument-invariant" benchmark that can be compared to other ATEs extrapolated from other instruments or alternative identification strategies.

Consider re-writing the structure given by Eqs (1)–(3), and D3, S8, and S9 as

$$\begin{aligned} Y_1 &= \mu_1 + U_1 \\ Y_0 &= \mu_0 + U_0 \\ D &= 1\left[Zg_1 + (1-Z)g_0 + U^D \geq 0\right] \\ Y &= DY_1 + (1-D)Y_0 \end{aligned}$$

where μ_1, μ_0 are constants, Z is a binary instrument, and (U_1, U_0, U^D) characterize both the individual's type and all other factors that determine Y and selection, and is independent of Z by D3 and S8. g_1 and g_0 are constants in the selection equation: S9 is satisfied. U_1, U_0, U^D can be normalized to be mean zero error terms. The ATE is by construction equal to $\mu_1 - \mu_0$.

Let us adopt the following functional form assumption:

- S10 (Normality of Errors): (U_1, U^D) and (U_0, U^D) are both bivariate normals with covariance matrices $\begin{pmatrix} \sigma_1 & \rho_1\sigma_1 \\ \rho_1\sigma_1 & 1 \end{pmatrix}$ and $\begin{pmatrix} \sigma_0 & \rho_0\sigma_0 \\ \rho_0\sigma_0 & 1 \end{pmatrix}$, respectively, where we are—without loss of generality—normalizing U^D to have a variance of 1.

This is simply the standard dummy endogenous variable system (as in Heckman (1976, 1978) and Maddala (1983)), with the special case of a dummy variable instrument, and is a case that is considered in recent work by Angrist (2004) and Oreopoulos (2006). With this one functional form assumption, we obtain a relationship between LATE and ATE.

In particular we know that

$$\text{LATE} = \mu_1 - \mu_0 + (\rho_1\sigma_1 - \rho_0\sigma_0)\frac{\phi(-g_1) - \phi(-g_0)}{\Phi(-g_0) - \Phi(-g_1)} \tag{10}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of the standard normal, respectively.[25] This is a standard result that directly follows from the early work on selection models (Heckman (1976, 1978). See also Heckman et al. (2001, 2003)). This framework has been used to discuss the relationship between ATE and LATE (see Heckman et al. (2001), Angrist (2004), and Oreopoulos (2006)). With a few exceptions (such as Heckman et al. (2001), for example), other applied researchers typically do not make use of the fact that even with information on just the three variables Y, D, and Z, the "selection correction" term in Eq. (10) can be computed. In particular

$$E[Y|D=1, Z=z] = \mu_1 + \rho_1\sigma_1\left(\frac{\phi(-g_z)}{1-\Phi(-g_z)}\right)$$
$$E[Y|D=0, Z=z] = \mu_0 - \rho_0\sigma_0\left(\frac{\phi(-g_z)}{\Phi(-g_z)}\right)$$

implies that

$$\rho_1\sigma_1 = \frac{E[Y|D=1,Z=1] - E[Y|D=1,Z=0]}{\frac{\phi(\Phi^{-1}(1-E[D|Z=1]))}{E[D|Z=1]} - \frac{\phi(\Phi^{-1}(1-E[D|Z=0]))}{E[D|Z=0]}} \tag{11}$$

and analogously that

$$\rho_0\sigma_0 = \frac{E[Y|D=0,Z=0] - E[Y|D=0,Z=1]}{\frac{\phi(\Phi^{-1}(1-E[D|Z=1]))}{1-E[D|Z=1]} - \frac{\phi(\Phi^{-1}(1-E[D|Z=0]))}{1-E[D|Z=0]}}. \tag{12}$$

Having identified $\rho_1\sigma_1$ and $\rho_0\sigma_0$, we have the expression

$$\text{ATE} = \text{LATE} - (\rho_1\sigma_1 - \rho_0\sigma_0) \cdot \frac{\phi\big(\Phi^{-1}(1-E[D|Z=1])\big) - \phi\big(\Phi^{-1}(1-E[D|Z=0])\big)}{E[D|Z=1] - E[D|Z=0]}. \tag{13}$$

[25] To see this, note that $E[Y|Z=1] = E[Y_1|D=1, Z=1]\Pr[D=1|Z=1] + E[Y_0|D=0, Z=1]\Pr[D=0|Z=1]$ which is equal to $(\mu_1 + \rho_1\sigma_1\frac{\phi(-g_1)}{1-\Phi(-g_1)})(1-\Phi(-g_1)) + (\mu_0 - \rho_0\sigma_0\frac{\phi(-g_1)}{\Phi(-g_1)})(\Phi(-g_1))$. We can decompose the first term into two terms to yield $(\mu_1 + \rho_1\sigma_1\frac{\phi(-g_0)}{1-\Phi(-g_0)})(1-\Phi(-g_0)) + (\mu_1(\Phi(-g_0) - \Phi(-g_1)) + \rho_1\sigma_1(\phi(-g_1) - \phi(-g_0))) + (\mu_0 - \rho_0\sigma_0\frac{\phi(-g_1)}{\Phi(-g_1)})(\Phi(-g_1))$. Taking the difference between this and an analogous expression for $E[Y|Z=0]$, the first and fourth terms cancel. Dividing the result by $\Phi(-g_0) - \Phi(-g_1)$ yields Eq. (10).

This expression is quite similar to that given in Section 3.1.4. Once again, the result of an ex post evaluation can be viewed as the leading term in the extrapolative goal of an ex ante evaluation: to obtain the effects of a program that was *not* implemented (i.e. random assignment with perfect compliance) or the ATE. This expression also shows how the goals of the ex post and ex ante evaluation problems can be complementary. Ex post evaluations aim to get the best estimate of the first term in the above equation, whereas ex ante evaluations are concerned with the assumptions necessary to extrapolate from LATE to ATE as in the above expression.

LATE versus ATE in practice

If S10 were adopted, how much might estimates of LATE differ from those of ATE in practice? To investigate this, we obtained data from a select group of empirical studies and computed both the estimates of LATE and the "selection correction" term in (13), using (11) and (12).[26]

The results are summarized in Table 1. For each of the studies, we present the simple difference in the means $E[Y|D = 1] - E[Y|D = 0]$, the Wald estimate of LATE, the implied selection error correlations, $\rho_1\sigma_1$, $\rho_0\sigma_0$, a selection correction term, and the implied ATE. For comparison, we also give the average value of Y. Standard deviations are in brackets. The next to last row of the table gives the value of the second term in (13), and its significance (calculated using the delta method).

The quantity $\rho_1\sigma_1 - \rho_0\sigma_0$ is the implied covariance between the gains $U_1 - U_0$ and the selection error U^D. A "selection on gains" phenomenon would imply $\rho_1\sigma_1 - \rho_0\sigma_0 > 0$. With the study of Abadie et al. (2002) in the first column, we see a substantial negative selection term, where the resulting LATE is significantly less than either the simple difference, or the implied ATE. If indeed the normality assumption is correct, this would imply that for the purposes of obtaining ATE, which might be the target of an ex ante evaluation, simply using LATE would be misleading, and ultimately even worse than using the simple difference in means.

On the other hand, in the analysis of Angrist and Evans (1998), the estimated LATE is actually quite similar to the implied ATE. So while a skeptic might consider it "uninteresting" to know the impact of having more than 2 children for the "compliers" (those whose family size was impacted by the gender mix of the first two children), it turns out in this context—if one accepts the functional form assumption—LATE and

[26] We chose the studies for this exercise in the following way. We searched articles mentioning "local average treatment effect(s)", as well as articles which cite Imbens and Angrist (1994) and Angrist et al. (1996). We restricted the search to articles that are published in *American Economic Review*, *Econometrica*, *Journal of Political Economy*, or *Quarterly Journal of Economics*, or *Review of Economic Studies*. From this group, we restricted our attention to studies in which both the instrument and the treatment were binary. The studies presented in the table are the ones in this group for which we were able to obtain the data, successfully replicate the results, and where computing the IV estimate without covariates did not substantially influence the results (Angrist and Evans, 1998; Abadie et al., 2002; Angrist et al., 2006; Field, 2007).

Table 1 Extrapolating from LATE to ATE

$$\text{ATE} = \text{LATE} - (\rho_1\sigma_1 - \rho_0\sigma_0) \cdot \underbrace{\frac{\phi(\Phi^{-1}(1 - E[D|Z=1])) - \phi(\Phi^{-1}(1 - E[D|Z=0]))}{E[D|Z=1] - E[D|Z=0]}}_{\star}$$

	Abadie et al. (2002)	**Angrist et al. (2002)**	**Angrist and Evans (1998)**	**Field (2007)**
Context	Effect of training on trainee earnings	Vouchers for private schooling in Colombia	Effect of childbearing on labor supply	Titling program for urban squatters in Peru
Outcome (Y)	Earnings	Schooling	Worked for pay	Total weekly work hours
Treatment (D)	Training	Scholarship use	More than two children	Title possession and report of experiencing change in tenure security
Instrument (Z)	Offer of training	Received a voucher	First two children are of the same sex	Program in neighborhood
(1) Mean of Y	19147	7.400	0.565	103.69
	[19540]	[1.099]	[0.496]	[77.68]
(2) $E[Y\|D=1] - E[Y\|D=0]$	3970	0.292	−0.121	−1.07
(3) LATE $= \frac{E[Y\|Z=1]-E[Y\|Z=0]}{E[D\|Z=1]-E[D\|Z=0]}$	1825	0.168	−0.132	31.59
(4) $\rho_1\sigma_1$	−3646	0.00012	0.0265	−27.01
(5) $\rho_0\sigma_0$	2287	0.326	−0.0088	−9.96
(6) $\rho_1\sigma_1 - \rho_0\sigma_0$: (4)–(5)	−5932	−0.326	0.035	−17.05
(7) $\star$ Term	0.57	−0.221	0.249	−0.30
(8) Selection correction term: −(6) × (7)	3400	−0.072	−0.0088	−5.07
	(882)$\star\star\star$	(0.032)$\star\star$	(0.0082)	(16.55)
(9) ATE: (3)+(8)	5225	0.096	−0.141	26.52

Standard deviations are in brackets. Standard errors (in parentheses) are calculated using the delta method. $\star\star$—Signicant at the 5% level, $\star\star\star$— Significant at the 1% level.

ATE do not differ very much.[27] The other studies are examples of intermediate cases: LATE may not be equal to ATE, but it is closer to ATE than the simple difference $E[Y|D = 1] - E[Y|D = 0]$.

Our point here is neither to recommend nor to discourage the use of this normal selection model for extrapolation. Rather, it is to illustrate and emphasize that even if LATE is identifying an average effect for a *conceptually* different population from the ATE, this does not necessarily mean that the two quantities in an actual application are very different. Our other point is that any inference that uses LATE to make *any* statement about the causal phenomena outside the context from which a LATE is generated, must necessarily rely on a structural assumption, whether implicitly or explicitly. In this discussion of extrapolating from LATE to ATE, we are being explicit that we are able to do this through a bivariate normal assumption. While such an assumption may seem unpalatable to some, it is clear that to insist on making no extrapolative assumptions is to abandon the ex ante evaluation goal entirely.

3.3. Regression discontinuity design: sharp

This section provides an extended discussion of identification and estimation of the regression discontinuity (RD) design. RD designs were first introduced by Thistlethwaite and Campbell (1960) as a way of estimating treatment effects in a non-experimental setting, where treatment is determined by whether an observed "assignment" variable (also referred to in the literature as the "forcing" variable or the "running" variable) exceeds a known cutoff point. In their initial application of RD designs, Thistlethwaite and Campbell (1960) analyzed the impact of merit awards on future academic outcomes, using the fact that the allocation of these awards was based on an observed test score. The main idea behind the research design was that individuals with scores just below the cutoff (who did not receive the award) were good comparisons to those just above the cutoff (who did receive the award). Although this evaluation strategy has been around for almost fifty years, it did not attract much attention in economics until relatively recently.

Since the late 1990s, a growing number of studies have relied on RD designs to estimate program effects in a wide variety of economic contexts. Like Thistlethwaite and Campbell (1960), early studies by Van der Klaauw (2002) and Angrist and Lavy (1999) exploited threshold rules often used by educational institutions to estimate the effect of financial aid and class size, respectively, on educational outcomes. Black (1999) exploited the presence of discontinuities at the geographical level (school district boundaries) to estimate the willingness to pay for good schools. Following these early papers in the area of education, the past five years have seen a rapidly growing literature using RD

[27] The obvious problem with the functional form here is that "Worked for pay" is a binary variable. One can still use the bivariate normal framework as an approximation, if Y is interpreted to be the latent probability of working, which can be continuously distributed (but one has to ignore the fact that the tails of the normal necessarily extend beyond the unit interval). In this case, the "average" effect is the average effect on the underlying probability of working.

designs to examine a range of questions. Examples include: the labor supply effect of welfare, unemployment insurance, and disability programs; the effects of Medicaid on health outcomes; the effect of remedial education programs on educational achievement; the empirical relevance of median voter models; and the effects of unionization on wages and employment.

An important impetus behind this recent flurry of research is a recognition, formalized by Hahn et al. (2001), that RD designs require seemingly mild assumptions compared to those needed for other non-experimental approaches. Another reason for the recent wave of research is the belief that the RD design is not "just another" evaluation strategy, and that causal inferences from RD designs are potentially more credible than those from typical "natural experiment" strategies (e.g. difference-in-difference or instrumental variables), which have been heavily employed in applied research in recent decades. This notion has a theoretical justification: Lee (2008) formally shows that one need not *assume* the RD design isolates treatment variation that is "as good as randomized"; instead, such randomized variation is a *consequence* of agents' inability to precisely control the assignment variable near the known cutoff.

So while the RD approach was initially thought to be "just another" program evaluation method with relatively little general applicability outside of a few specific problems, recent work in economics has shown quite the opposite.[28] In addition to providing a highly credible and transparent way of estimating program effects, RD designs can be used in a wide variety of contexts covering a large number of important economic questions. These two facts likely explain why the RD approach is rapidly becoming a major element in the toolkit of empirical economists.

Before presenting a more formal discussion of various identification and estimation issues, we first briefly highlight what we believe to be the most important points that have emerged from the recent theoretical and empirical literature on the RD design.[29] In this chapter, we will use V to denote the assignment variable, and treatment will be assigned to individuals when V exceeds a known threshold c, which we later normalize to 0 in our discussion.

- **RD designs can be invalid if individuals can precisely manipulate the "assignment variable".**
  When there is a payoff or benefit to receiving a treatment, it is natural for an economist to consider how an individual may behave to obtain such benefits. For example, if students could effectively "choose" their test score V through effort, those who chose a score c (and hence received the merit award) could be somewhat different from

[28] See Cook (2008) for an interesting history of the RD design in education research, psychology, statistics, and economics. Cook argues the resurgence of the RD design in economics is unique as it is still rarely used in other disciplines.

[29] Recent surveys of the RD design in theory and practice include Lee and Lemieux (2009), Van der Klaauw (2008a), and Imbens and Lemieux (2008a).

those who chose scores just below c. The important lesson here is that the existence of a treatment being a discontinuous function of an assignment variable is *not* sufficient to justify the validity of an RD design. Indeed, if anything, discontinuous rules may generate incentives, causing behavior that would *invalidate* the RD approach.

- **If individuals—even while having some influence—are unable to *precisely* manipulate the assignment variable, *a consequence* of this is that the variation in treatment near the threshold is randomized as though from a randomized experiment.**
  This is a crucial feature of the RD design, and a reason that RD designs are often so compelling. Intuitively, when individuals have imprecise control over the assignment variable, even if some are especially likely to have values of V *near* the cutoff, *every* individual will have approximately the same probability of having an V that is just above (receiving the treatment) or just below (being denied the treatment) the cutoff—similar to a coin-flip experiment. This result clearly differentiates the RD and IV (with a non-randomized instrument) approaches. When using IV for causal inference, one must *assume* the instrument is exogenously generated as if by a coin-flip. Such an assumption is often difficult to justify (except when an actual lottery was run, as in Hearst et al. (1986) or Angrist (1990), or if there were some biological process, e.g. gender determination of a baby, mimicking a coin-flip). By contrast, the variation that RD designs isolate is randomized *as a consequence* of the assumption that individuals have imprecise control over the assignment variable (Lee, 2008).
- **RD designs can be analyzed—and tested—like randomized experiments.**
  This is the key implication of the local randomization result. If variation in the treatment near the threshold is approximately randomized, then it follows that all "baseline characteristics" – all those variables determined prior to the realization of the assignment variable—should have the same distribution just above and just below the cutoff. If there is a discontinuity in these baseline covariates, then at a minimum, the underlying identifying assumption of individuals' inability to precisely manipulate the assignment variable is unwarranted. Thus, the baseline covariates are used to *test* the validity of the RD design. By contrast, when employing an IV or a matching/regression-control strategy in non-experimental situations, assumptions typically need to be made about the relationship of these other covariates to the treatment and outcome variables.[30]
- **The treatment effects from RD can be interpreted as a weighted average treatment effect.**
  It is tempting to conclude that the RD delivers treatment effects that "only apply" for the sub-population of individuals whose V is arbitrarily close to the threshold c. Such an interpretation would imply that the RD identifies treatment effects for "virtually

[30] Typically, one assumes that *conditional on the covariates*, the treatment (or instrument) is "as good as" randomly assigned.

no one". Fortunately, as we shall see below, there is an alternative interpretation: the average effect identified by a valid RD is that of a weighted average treatment effect where the weights are the relative ex ante probability that the value of an individual's assignment variable will be in the neighborhood of the threshold (Lee, 2008).

Randomized experiments from non-random selection

As argued in Lee and Lemieux (2009), while there are some mechanical similarities between the RD design and a "matching on observables" approach or between the RD design and an instrumental variables approach, the RD design can instead be viewed as a close "cousin" of the randomized experiment, in the sense that what motivates the design and what "dictates" the modeling is specific institutional knowledge of the treatment assignment process. We illustrate this by once again using the common framework given by Eqs (1)–(3). We begin with the case of the "sharp" RD design, whereby the treatment status is a deterministic "step-function" of an observed assignment variable V. That is, $D = 1$ if and only if V crosses the discontinuity threshold (normalized to 0 in our discussion).

Returning to our hypothetical job search assistance program, suppose that the state agency needed to ration the number of participants in the program, and therefore mandated treatment (personal visit and/or phone call from job counselor) for those whom the agency believed would the program would most greatly benefit. In particular, suppose the agency used information on individuals' past earnings and employment information to generate a score V that indicated the likely benefit of the program to the individual, and determined treatment status based on whether that V exceeded the threshold 0.

Such a discontinuous rule can be described as

- D4: (Discontinuous rule) $D = 1[V \geq 0]$: This implies that Eq. (2) becomes $P^* = p^*(W, U) = \Pr[V \geq 0|W, U]$.
- D5: V is observed.

Both D4 and D5 come from institutional knowledge of how treatment is assigned, and thus are more descriptions ("D") than assumptions.

We further assume that

- S11: (Positive density at the threshold): $f_V(0) > 0$.

This assumption ensures that there are some individuals at the threshold.

Since we have introduced a variable V that is realized before D, we must specify its relation to Y, so that Eq. (3) becomes

- S12 (Continuous impact of V) $Y = y(D, W, U, V)$, where $y(d, w, u, v)$ is continuous in v (at least in a neighborhood of $v = 0$).

This assumption states that for any individual type U, as V crosses the discontinuity threshold 0, any change in Y must be attributable to D and D only. As we shall see,

this assumption, while necessary, is *not* sufficient for the RD to deliver valid causal inferences.

The most important assumption for identification is

- S13 (Continuous density; incomplete/imprecise control of V) The distribution of V conditional on U has density $f_{V|U}(v)$ that is continuous in v, at least in the neighborhood of $v = 0$.

This condition, which we will discuss in greater detail below, says that individuals—no matter how much they can influence the *distribution* of V with their actions—cannot *precisely* control V, even if they may make decisions W in anticipation of this uncertainty.

There are at least two alternative interpretations of this condition. One is that individuals may actually precisely control V, but they are responding to different external factors, which generates a distribution of different possible Vs that could occur depending on these outside forces. S13 says that the distribution of V—as driven by those outside forces—must have a continuous density. The other interpretation is that even conditional on U, there exists heterogeneity in the factors that determine V. In this case, S13 is a statement about the distribution of this heterogeneity in V having continuous density conditional on the type U.

As Lee (2008) shows, it is precisely S13 that will generate a local randomization result. In particular, S13 implies

$$\begin{aligned}\lim_{v\uparrow 0} f_{U|V=v}(u) &= \lim_{v\uparrow 0} \frac{f_{V|U=u}(v)}{f_V(v)} \cdot f_U(u) \\ &= \frac{f_{V|U=u}(0)}{f_V(0)} \cdot f_U(u) \\ &= f_{U|V=0}(u)\end{aligned}$$

which says that the distribution of the unobserved "types" U will be approximately equal on either side of the discontinuity threshold in a neighborhood of 0. This is the sense in which it accomplishes *local randomization*, akin to the randomization in an experiment. The difference is in how the problem expressed in Eq. (5) is being confronted. The experimenter in the randomized experiment ensures that treated and non-treated individuals have the same distribution of latent propensities P^* by dictating that all individuals have the same fixed $P^* = p_0$. In the non-experimental context here, we have no control over the distribution of P^*, but if S13 holds, then the (non-degenerate) distribution of P^* (which is a function of U) will approximately be equal between treated and non-treated individuals—for those with realized V in a small neighborhood of 0.

Now consider the expectation of Y at the discontinuity threshold

$$E[Y|V = 0] = E[y(1, w(U), U, 0)|V = 0]$$

$$= \int y(1, w(u), u, 0) f_{U|V=0}(u) \mathrm{d}u$$
$$= \int y(1, w(u), u, 0) \frac{f_{V|U=u}(0)}{f_V(0)} \cdot f_U(u) \mathrm{d}u$$

where the third line follows from Bayes' Rule. Similarly, we have

$$\lim_{v \uparrow 0} E[Y|V = v] = \lim_{v \uparrow 0} E[y(0, w(U), U, v)|V = v]$$
$$= \lim_{v \uparrow 0} \int y(0, w(u), u, v) \frac{f_{V|U=u}(v)}{f_V(v)} \cdot f_U(u) \mathrm{d}u$$
$$= \int y(0, w(u), u, 0) \frac{f_{V|U=u}(0)}{f_V(0)} \cdot f_U(u) \mathrm{d}u$$

where the last line follows from the continuity assumption S12 and S13.

The RD estimand—the difference between the above two quantities—is thus

$$E[Y|V = 0] - \lim_{v \uparrow 0} E[Y|V = v] = \int [y(1, w(u), u, 0) - y(0, w(u), u, 0)] \times \frac{f_{V|U=u}(0)}{f_V(0)} \cdot f_U(u) \mathrm{d}u$$
$$= \int \Delta(w(u), u, 0) \frac{f_{V|U=u}(0)}{f_V(0)} \cdot f_U(u) \mathrm{d}u. \quad (14)$$

That is, the discontinuity in the conditional expectation function $E[Y|V = v]$ identifies a *weighted* average of causal impacts $\Delta(w(u), u, 0)$, where the weights are proportional to type U's relative likelihood that the assignment variable V is in the neighborhood of the discontinuity threshold 0.

Equation (14) provides a quite different alternative to the interpretation of the estimand as "the treatment effect for those whose V are close to zero"—which connotes a very limited inference, because in the limit, there are no individuals at $V = 0$. The variability in weights in (14) depend very much on the typical scale of $f_{V|U}(\cdot)$ relative to the location of $f_{V|U}(\cdot)$ across the U types. That is, if for each type U there is negligible variability in V, then the RD estimand will indeed identify a treatment effect for those individuals who can be most expected to have V close to the threshold. On the other extreme, if there is large variability, with $f_{V|U}(\cdot)$ having flat tails, the weights will tend to be more uniform.[31]

One of the reasons why RD design can be viewed as a "cousin" of the randomized experiment is that the latter is really a special case of the sharp RD design.

[31] For example, for the uniform density $f_{V|U}(v) = \frac{1}{\theta} \cdot 1[\mu_U \leq v \leq \mu_U + \theta]$, the weights would be identical across types, even though there would be variability in the probability of treatment driven by variability in μ_U.

When randomly assigning treatment, one can imagine accomplishing this through a continuously distributed random variable V that has the same distribution for every individual. In that case, V would not enter the function determining Y, and hence S12 would be unnecessary. It would follow that $f_{V|U}(v) = f_V(v)$, and consequently every individual would receive equal weight in the average effect expression in Eq. (14).

The final point to notice about the average effect that is identified by the RD design is that it is a weighted average of $\Delta(w(U), U, 0)$, which may be different from a weighted average of $\Delta(w(U), U, v)$ for some other v, even if the weights $\frac{f_{V|U=u}(0)}{f_V(0)}$ do not change. Of course, there is no difference between the two quantities in situations where V is thought to have no impact on the outcome (or the individual treatment effect). For example, if a test score V is used for one and only one purpose—to award a scholarship D—then it might be reasonable to assume that V has no other impact on future educational outcomes Y. As discussed in Lee (2008) and Lee and Lemieux (2009), in other situations, the concept of $\Delta(w(U), U, v)$ for values of v other than 0 may not make much practical sense. For example, in the context of estimating the electoral advantage to incumbency in US House elections (Lee, 2008), it is difficult to conceive of the counterfactual $y(0, w(u), u, v)$ when v is away from the threshold: what does it mean for the outcome that *would* have been obtained if the candidate who became the incumbent with 90 percent of the vote had *not* become the incumbent, having won 90 percent of the vote? Here, incumbent status is defined by V.

3.3.1. Assessment: Valid or invalid RD?

We now assess this design on the basis of the three criteria discussed in Section 2.2.1. First, some of the conditions for identification are indeed literally descriptions of the assignment process: D4 and D5. Others, like S11 through S13, are conjectures about the assignment process, and the underlying determinants of Y. S11 requires that there is positive density of V at the threshold. S12 allows V—which can be viewed as capturing "all other factors" that determine D—to have its own structural effect on Y. It is therefore not as restrictive as a standard exclusion restriction, but S12 does require that the impact of V is continuous. S13 is the most important condition for identification, and we discuss it further below.

Our second criterion is the extent to which inferences could be consistent with many competing behavioral models. The question is to what extent does S12 and S13 restrict the class of models for which the RD causal inference remains valid? When program status is determined solely on the basis of a score V, and V is used for nothing else but the determination of D, we expect most economic models to predict that the only reason why V would have a discontinuous impact on Y would be because an individual's status switches from non-treated to treated. So, from the perspective of modeling economic behavior, S12 does not seem to be particularly restrictive.

By contrast, S13 is potentially restrictive, ruling out some plausible economic behavior. Are individuals able to influence the assignment variable, and if so, what is the

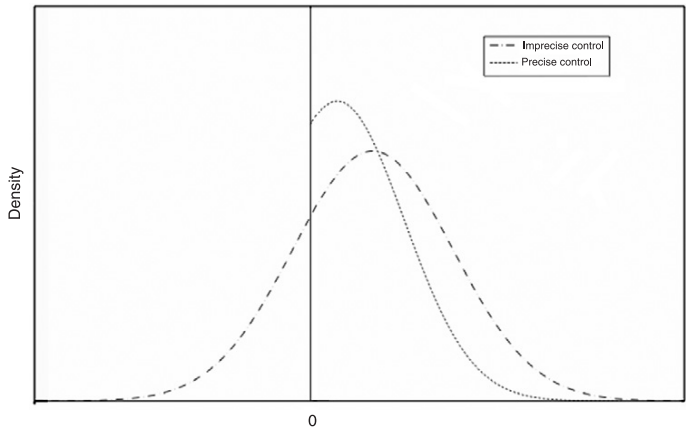


Figure 1 ***Density of Assignment Variable Conditional on*** $W = w, U = u$.

nature of this control? This is probably the most important question to ask when assessing whether a particular application should be analyzed as an RD design. If individuals have a great deal of control over the assignment variable and if there is a perceived benefit to a treatment, one would certainly expect individuals on one side of the threshold to be systematically different from those on the other side.

Consider the test-taking example from Thistlethwaite and Campbell (1960). Suppose there are two types of students: A and B. Suppose type A students are more able than B types, and that A types are also keenly aware that passing the relevant threshold (50 percent) will give them a scholarship benefit, while B types are completely ignorant of the scholarship and the rule. Now suppose that 50 percent of the questions are trivial to answer correctly, but due to random chance, students will sometimes make careless errors when they initially answer the test questions, but would certainly correct the errors if they checked their work. In this scenario, only type A students will make sure to check their answers before turning in the exam, thereby assuring themselves of a passing score. The density of their score is depicted in the truncated density in Fig. 1. Thus, while we would expect those who barely passed the exam to be a mixture of type A and type B students, those who barely failed would exclusively be type B students. In this example, it is clear that the marginal failing students do *not* represent a valid counterfactual for the marginal passing students. Analyzing this scenario within an RD framework would be inappropriate.

On the other hand, consider the same scenario, except assume that questions on the exam are *not* trivial; there are no guaranteed passes, no matter how many times the students check their answers before turning in the exam. In this case, it seems more plausible that among those scoring near the threshold, it is a matter of "luck" as to which side of the threshold they land. Type A students can exert more effort—because

they know a scholarship is at stake—but they do not know the exact score they will obtain. This can be depicted by the untruncated density in Fig. 1. In this scenario, it would be reasonable to argue that those who marginally failed and passed would be otherwise comparable, and that an RD analysis *would* be appropriate and would yield credible estimates of the impact of the scholarship.

These two examples make it clear that one must have some knowledge about the mechanism generating the assignment variable, beyond knowing that if it crosses the threshold, the treatment is "turned on". The"folk wisdom" in the literature is to judge whether the RD is appropriate based on whether individuals could *manipulate* the assignment variable and *precisely* "sort" around the discontinuity threshold. The key word here should be "precise", rather than "manipulate". After all, in both examples above, individuals do exert some control over the test score. And indeed in virtually every known application of the RD design, it is easy to tell a plausible story that the assignment variable is to some degree influenced by *someone*. But individuals will not always have *precise* control over the assignment variable. It should, perhaps, seem obvious that it is necessary to rule out *precise* sorting to justify the use of an RD design. After all, individual self-selection into treatment or control regimes is exactly why simple comparison of means is unlikely to yield valid causal inferences. Precise sorting around the threshold is self-selection.

Finally, the data generating process given by D4, D5, S11, S12, and S13 has many testable implications. D4 and D5 are directly verifiable, and S11 can be checked since the marginal density $f_V(0)$ can be observed from the data. S12 appears fundamentally unverifiable, but S13, which generates the local randomization result, is testable in two ways. First, as McCrary (2008) points out, the continuity of each type's density $f_{V|U}(\cdot)$ (S13) implies that the observed marginal density $f_V(\cdot)$ is continuous, leading to a natural test: examining the data for a potential discontinuity in the density of V at the threshold; McCrary (2008) proposes an estimator for this test. Second, the local randomization result implies that

$$\lim_{v \uparrow 0} F_{W|V=v}(w) = F_{W|V=0}(w)$$

because W is by definition determined prior to D (Eq. (1)). This is analogous to the test of randomization where treated and control observations are tested for balance in observable elements of W.[32]

In summary, based on the above three criteria, the RD design does appear to have potential to deliver highly credible causal inferences because some important aspects of the model are literal descriptions of the assignment process and because the conditions

[32] Additionally, *any* variable determined prior to D—whether or not they are an element of W—should have the same distribution on either side of the discontinuity threshold. This, too, is analogous to the case of randomized assignment.

for identification are consistent with a broad class of behavioral models (e.g. V can be endogenous—as it is influenced by actions in W—as long there is "imprecise" manipulation of V). Perhaps most importantly, the key condition that is *not* derived from our institutional knowledge (and is the key restriction on economic behavior (S13)), has a testable implication that is as strong as that given by the randomized experiment.

3.3.2. *Ex ante evaluation: extrapolating from the RD to treatment on the treated*

Consider again our hypothetical job search assistance program, where individuals were assigned to the treatment group if their score V (based on earnings and employment information and the state agency's model of who would most benefit from the program) exceeded the threshold 0. From an ex ante evaluation perspective, one could potentially be interested in predicting the policy impact of "shutting down the program" Heckman and Vytlacil (2005). As pointed out in Heckman and Vytlacil (2005), the treatment on the treated parameter (TOT) is an important ingredient for such a prediction. But as we made precise in the ex post evaluation discussion, the RD estimand identifies a particular weighted average treatment effect that in general will be different from the TOT. Is there a way to extrapolate from the average effect in (14) to TOT?

We now sketch out one such proposal for doing this, recognizing this is a relatively uncharted area of research (and that TOT, while an ingredient in computing the impact of the policy of "shutting down the program", may not be of interest for a different proposed policy). Using D4, D5, S11, S12, and S13, we can see that the TOT is

$$\begin{aligned} E[\Delta(w(U), U, V)|D = 1] &= E[y(1, w(U), U, V) - y(0, w(U), U, V)|D = 1] \\ &= E[Y|D = 1] - E[y(0, w(U), U, V)|V > 0]. \end{aligned}$$

This reveals that the key missing ingredient is the second term of this difference.

Note that we have

$$\begin{aligned} &E[y(0, w(U), U, V)|V \geq 0] \\ &= \frac{1}{1 - F_V(0)} \int_{v \geq 0} \left[\int y(0, w(u), u, v) \mathrm{d}F_{U|V=v}(u) \right] f_V(v) \mathrm{d}v \\ &= \frac{1}{1 - F_V(0)} \int_{v \geq 0} \left[\int y(0, w(u), u, v) \frac{f_{V|U=u}(v)}{f_V(v)} \mathrm{d}F_U(u) \right] f_V(v) \mathrm{d}v \end{aligned}$$

where the second line again follows from Bayes' rule.

Suppose V has bounded support, and also assume

- S14: (Differentiable Density) Let $f_{V|U=u}(v)$ have continuous qth derivative on $v \geq 0$, for all u.
- S15: (Differentiable Outcome Function) Let $y(0, w(u), u, v)$ have continuous qth derivative on $v \geq 0$, for all u.

With the addition of S14 and S15, this will imply that we can use the Taylor approximation

$$\begin{aligned}
E[y(0, w(U), U, v)|V = v] \approx\ & E[y(0, w(U), U, 0)|V = 0] \\
& + \left.\frac{\partial E[y(0, w(U), U, v^*)|V = v^*]}{\partial v^*}\right|_{v^*=0} v + \cdots \\
& + \frac{1}{q!}\left.\frac{\partial^q E[y(0, w(U), U, v^*)|V = v^*]}{\partial v^{*q}}\right|_{v^*=0} v^q \\
\approx\ & \lim_{v^*\uparrow 0} E[Y|V = v^*] + \lim_{v^*\uparrow 0}\left.\frac{\partial E[Y|V = v^*]}{\partial v^*}\right|_{v^*=0} v \\
& + \cdots + \frac{1}{q!}\lim_{v^*\uparrow 0}\left.\frac{\partial^q E[Y|V = v^*]}{\partial v^{*q}}\right|_{v^*=0} v^q.
\end{aligned}$$

In principle, once this function can be (approximately) identified, one can average the effects over the treated population using the conditional density $f_{V|V\geq 0}(v)$.

Once again, we see that the leading term of an extrapolation for an ex ante evaluation, is related to the results of an ex post evaluation: it is precisely the counterfactual average in Eq. (14). There are a number of ways of estimating derivatives nonparametrically (Fan and Gijbels, 1996; Pagan and Ullah, 1999), or one could alternatively attempt to approximate the function $E[Y|V = v]$ for $D < 0$ with a low-order global polynomial. We recognize that, in practical empirical applications, estimates of higher order derivatives may be quite imprecise.

Nevertheless, we provide this simple method to illustrate the separate role of conditions for credible causal inference (D4, D5, S11, S12, and S13), and the additional structure needed (S14 and S15) to identify a policy-relevant parameter such as the TOT. What alternative additional structure could be imposed on Eqs (1)–(3) to identify parameters such as the TOT seems to be an open question, and likely to be somewhat context-dependent.

3.3.3. Estimation issues for the RD design

There has been a recent explosion of applied research utilizing the RD design (see Lee and Lemieux (2009)). From this applied literature, a certain "folk wisdom" has emerged about sensible approaches to implementing the RD design in practice. The key challenge with the RD design is how to use a finite sample of data on Y and V to estimate the conditional expectations in the discontinuity $E[Y|V = 0] - \lim_{v\uparrow 0} E[Y|V = v]$. Lee and Lemieux (2009) discuss these common practices and their justification in greater detail. Here we simply highlight some of the key points of that review, and then conclude with the recommended "checklist" suggested by Lee and Lemieux (2009).[33]

[33] See Imbens and Lemieux (2008b) and Van der Klaauw (2008b) for other surveys.

- **Graphical presentation of an RD design is helpful and informative, but the visual presentation should not be tilted toward either finding an effect or finding no effect.**
  It has become standard to summarize RD analyses with a simple graph showing the relationship between the outcome and assignment variable. This has several advantages. The presentation of the "raw data" enhances the transparency of the research design. A graph can also give the reader a sense of whether the "jump" in the outcome variable at the cutoff is unusually large compared to the bumps in the regression curve away from the cutoff. Also, a graphical analysis can help identify why different functional forms give different answers, and can help identify outliers, which can be a problem in any empirical analysis. The problem with graphical presentations, however, is that there is some room for the researcher to construct graphs making it seem as though there are effects when there are none, or hiding effects that truly exist. A way to guard against this visual bias is to partition V into intervals—with the discontinuity threshold at one of the boundaries—and present the mean within each interval. Often, the data on V will already be discrete. This way, the behavior of the function around the threshold is given no special "privilege" in the presentation, yet it allows for the data to "speak for itself" as to whether there is an important jump at the threshold.[34]
- **Non-parametric estimation does not represent a unique, always-preferred "solution" to functional form issues raised by RD designs. It is therefore helpful to view it as a complement to—rather than a substitute for—parametric estimation.**
  Here it is helpful to keep distinct the notions of identification and estimation. The RD design, as discussed above, is non-parametrically identified, and no parametric restrictions are needed to compute $E[Y|V = 0] - \lim_{v \uparrow 0} E[Y|V = v]$, given an infinite amount of data. But with a finite sample, one has a choice of different statistics, some referred to as "non-parametric" (e.g. kernel regression, local linear regression), while others considered "parametric" (e.g. a low-order polynomial). As Powell (1994) points out, it is perhaps more helpful to view *models* rather than particular statistics as "parametric" or "non-parametric".[35] The bottom line is that when the analyst chooses a particular functional form (say, a low-order polynomial) in estimation, and the true function does not belong to that polynomial class, then the resulting estimator will, in general, be biased. When the analyst uses a non-parametric procedure such as local linear regression—essentially running a regression using only data points "close" to the cutoff—there will also be bias.[36] With a finite sample, it is impossible

[34] See Lee and Card (2008) for a discussion.

[35] As an example, Powell (1994) points out that the same least squares estimator can simultaneously be viewed as solutions to parametric, semi-parametric, and nonparametric problems.

[36] Unless the underlying function is exactly linear in the area being examined.

to know which case has a smaller bias without knowing something about the true function. There will be some functions where a low-order polynomial is a very good approximation and produces little or no bias, and therefore it is efficient to use all data points—both "close to" and "far away" from the threshold. In other situations, a polynomial may be a bad approximation, and smaller biases will occur with a local linear regression.

In practice, parametric and non-parametric approaches lead to the computation of the exact same statistic. For example, the procedure of regressing the outcome Y on V and a treatment dummy D and an interaction $V \cdot D$, can be viewed as a "parametric" regression *or* a local linear regression with a very large bandwidth. Similarly, if one wanted to exclude the influence of data points in the tails of the X distribution, one could call the exact same procedure "parametric" after trimming the tails, or "non-parametric" by viewing the restriction in the range of X as a result of using a smaller bandwidth.[37] Our main suggestion in estimation is to not rely on one particular method or specification. In any empirical analysis, results that are stable across alternative and equally plausible specifications are generally viewed as more reliable than those that are sensitive to minor changes in specification. RD is no exception in this regard.

- **Goodness-of-fit and other statistical tests can help rule out overly restrictive specifications.**

  Often the consequence of trying many different specifications is that it may result in a wide range of estimates. Although there is no simple formula that works in all situations and contexts for weeding out inappropriate specifications, it seems reasonable, at a minimum, not to rely on an estimate resulting from a specification that can be rejected by the data when tested against a strictly more flexible specification. For example, it seems wise to place less confidence in results from a low-order polynomial model, when it is rejected in favor of a less restrictive model (e.g., separate means for each discrete value of V). Similarly, there seems little reason to prefer a specification that uses all the data, if using the same specification but restricting to observations closer to the threshold gives a substantially (and statistically) different answer.

[37] One of the reasons why typical non-parametric "methods" (e.g. local linear regression) are sometimes viewed as being superior is that the statistics yield consistent estimators. But it is important to remember that such consistency is arising from an asymptotic approximation that dictates that one of the "parameters" of the statistic (i.e. the function of the sample data)—the bandwidth—shrinks (at an appropriate rate) as the sample size increases. Thus, the consistency of the estimator is a direct result of a different notion of asymptotic behavior. If one compares the behavior of "non-parametric" statistics (e.g. local linear regression) with that of "parametric" statistics (e.g. global polynomial regression) using the same asymptotic framework (i.e. statistics are not allowed to change with the sample size), then the non-parametric method loses this superiority in terms of consistency. Depending on the true underlying function (which is unknown), the difference between the truth and the probability limit of the estimator, may be larger or smaller with the "parametric" statistic.

A Recommended "checklist" for implementation

Below we summarize the recommendations given by Lee and Lemieux (2009) for the analysis, presentation, and estimation of RD designs.

1. **To assess the possibility of manipulation of the assignment variable, show its distribution.** The most straightforward thing to do is to present a histogram of the assignment variable, after partitioning the support of V into intervals; in practice, V may have a natural discreteness to it. The bin widths should be as small as possible, without compromising the ability to visually see the overall shape of the distribution. The bin-to-bin jumps in the frequencies can provide a sense in which any jump at the threshold is "unusual". For this reason, we recommend *against* plotting a smooth function comprised of kernel density estimates. A more formal test of a discontinuity in the density can be found in McCrary (2008).
2. **Present the main RD graph using binned local averages.** As with the histogram, the recommendation is to graphically present the sample means within the defined intervals. The non-overlapping nature of the bins for the local averages is important; we recommend against simply presenting a continuum of nonparametric estimates (with a single break at the threshold), as this will naturally tend to give the impression of a discontinuity even if there does not exist one in the population. Lee and Lemieux (2009) suggest a cross-validation procedure as well as simple ways to test the bin width choice against less restrictive alternatives. They recommend generally "undersmoothing", while at the same time avoiding "too narrow" bins that produce a scatter of data points, from which it is difficult to see the shape of the underlying function. Indeed, they also recommend against simply plotting the raw data without a minimal amount of local averaging.
3. **Graph a benchmark polynomial specification.** Super-impose onto the above graph the predicted values from a low-order polynomial specification. One can often informally assess, by comparing the two functions, whether a simple polynomial specification is an adequate summary of the data. In a way, these two functions give a sense of the range of functions that would fit the data. On the one hand, the local averages represent a flexible "non-parametric" representation of the true underlying function. On the other hand, a polynomial represents a "best case" scenario in terms of the variance of the RD estimate: if the true function really is a polynomial of the chosen order, standard regression theory suggests that the least squares estimator (that uses all the observations) will be unbiased, and potentially efficient in a class of all linear unbiased estimators.
4. **Explore the sensitivity of the results to a range of bandwidths, and a range of orders to the polynomial.** Lee and Lemieux (2009) provide an example of how to systematically examine different degrees of smoothing, through different bandwidths or polynomial orders, using both cross-validation to provide a rough guide to sensible

bandwidths, and the Akaike Information Criterion (AIC) as a rough guide to sensible orders for the polynomial. A useful graphical device for illustrating the sensitivity of the results to bandwidths is to plot the local linear discontinuity estimate against a continuum of bandwidths (within a range of bandwidths that are not ruled out by available specification tests). For an example of such a presentation, see the online appendix to Card et al. (2009a) and Lee and Lemieux (2009).

5. **Conduct a parallel RD analysis on the baseline covariates.** As discussed earlier, if the assumption that there is no precise manipulation or sorting of the assignment variable is valid, then there should be no discontinuities in variables that are determined prior to the assignment.
6. **Explore the sensitivity of the results to the inclusion of baseline covariates.** As discussed above, in a neighborhood of the discontinuity threshold, pre-determined covariates will have the same distribution on either side of the threshold, implying that inclusion of those covariates in a local regression should not affect the estimated discontinuity. If the estimates do change in an important way, it may indicate a potential sorting of the assignment variable that may be reflected in a discontinuity in one or more of the baseline covariates. Lee and Lemieux (2009) show how the assumption that the covariates can be approximated by the same order of polynomial as Y as a function of V can be used to justify including the covariates linearly in a polynomial regression.

Although it is impractical for researchers to present every permutation of presentation (e.g. points 2-4 for every one of 20 baseline covariates), probing the sensitivity of the results to this array of tests and alternative specifications—even if they only appear in online appendices—is an important component of a thorough RD analysis.

3.4. Regression discontinuity design: fuzzy

Returning to our hypothetical job search assistance program, consider the same setup as the RD design described above: based on past employment and earnings and a model of who would most benefit from the program, the government constructs a score V where those with $V \geq 0$ will receive the treatment (phone call/personal visit of job counselor). But now assume that V crossing the threshold 0 only determines whether the agency explicitly provides information to the individual about the existence of the program. In other words, V determines Z—as defined in the case of random assignment with imperfect compliance, discussed above in Section 3.2. As a result, D is no longer a deterministic function of V, but $\Pr[D = 1|V]$ is a potentially discontinuous function in V. This is known as the "fuzzy" RD design (see Hahn et al. (2001), for example, for a formal definition).

The easiest way to understand the fuzzy RD design is to keep in mind that the relationship between the fuzzy design and the sharp design parallels the relation between

random assignment with imperfect compliance and random assignment with perfect compliance. Because of this parallel, our assessment of the potential internal validity of the design and potential caveats follows the discussion in Section 3.2. Testing the design follows the same principles as in Section 3.2. Furthermore, one could, in principle, combine the extrapolative ideas in Section 3.3.2 and Section 3.2.2 to use fuzzy RD design estimates to make extrapolations along two dimensions: predicting the effect under mandatory compliance, and predicting the average effects at points away from the threshold.

We limit our discussion here to making explicit the conditions for identification of average effects within the common econometric framework we have been utilizing. The different conditions for this case are

- D6: (Discontinuous rule for Z) $Z = 1[V \geq 0]$. This implies that P^* in Eq. (2) is given by the function $p^*(W, U) = \Pr[V \geq 0|W, U]p_1^*(W, U) + \Pr[V < 0|W, U]p_0^*(W, U)$, where $p_z^*(W, U) \equiv \Pr[D = 1|W, U, Z = z]$.
- S14: (Exclusion Restriction) $Y = y(D, W, U, V)$, where $y(d, x, w, v)$ is continuous in v (at least in a neighborhood of $v = 0$). $W = w(U, V)$, where $w(u, v)$ is continuous in v (at least in a neighborhood of $v = 0$). (Z does not enter either function).

In sum, we have (1) the conditions from the Sharp RD (D5 (V observable), S11 (Positive Density at Threshold), S13 (Continuous Density at Threshold)), (2) a condition from the randomized experiment with imperfect compliance (S9 (Probabilistic Monotonicity)), and (3) two new "hybrid" conditions—D6 (Discontinuous rule for Z) and S14 (Exclusion Restriction).

Given the discontinuous rule D6, we have

$$E[Y|V = 0] = \int y(0, w(u), u, 0) + p_1^*(w(u), u)\Delta(w(u), u, 0)\mathrm{d}F_{U|V=0}(u)$$
$$E[Y|V = v] = \int y(0, w(u), u, v) + p_0^*(w(u), u)\Delta(w(u), u, v)\mathrm{d}F_{U|V=v}(u)$$

and because of S11, S13, and S14, we can combine the integrals so that the difference equals

$$E[Y|V = 0] - \lim_{v \uparrow 0} E[Y|V = v] = \int \Delta(w(u), u, 0)(p_1^*(w(u), u) - p_0^*(w(u), u))\frac{f_{V|U=u}(0)}{f_V(0)}\mathrm{d}F_U(u).$$

Normalizing the difference by the quantity $E[D|V=0] - \lim_{v\uparrow 0} E[D|V=v]$ yields

$$\frac{E[Y|V=0] - \lim_{v\uparrow 0} E[Y|V=v]}{E[D|V=0] - \lim_{v\uparrow 0} E[D|V=v]} = \int \Delta(w(u), u, 0) \times \left[\frac{(p_1^*(w(u), u) - p_0^*(w(u), u))}{E[D|V=0] - \lim_{v\uparrow 0} E[D|V=v]} \frac{f_{V|U=u}(0)}{f_V(0)} \right] \mathrm{d}F_U(u)$$

where the normalizing factor ensures that the weights in square brackets average to one.[38]

Thus the fuzzy RD estimand is a weighted average of treatment effects. The weights reflect two factors: the relative likelihood that a given type U's V will be close to the threshold reflected in the term $\frac{f_{V|U=u}(0)}{f_V(0)}$, and the influence that V crossing the threshold has on the probability of treatment, as reflected in $p_1^*(w(u), u) - p_0^*(w(u), u)$. S9 ensures these weights are nonnegative.

From a purely ex ante evaluation perspective, the weights would seem peculiar, and not related to any meaningful economic concept. But from a purely ex post evaluation perspective, the weights are a statement of fact. As soon as one believes that there is causal information in comparing Y just above and below the threshold—and such an intuition is entirely driven by the institutional knowledge given by D5 and D6—then it appears that the *only* way to obtain *some* kind of average effect (while remaining as agnostic as possible about the other unobservable mechanisms that enter the latent propensity P^*) with the data Y, V, D, is to make sure that the implied weights integrate to 1. We have no choice but to divide the difference by $E[D|V=0] - \lim_{v\uparrow 0} E[D|V=v]$.

As we have argued in previous sections, rather than abandon potentially highly credible causal evidence because the data and circumstance that we are handed did not deliver us the "desired" weights, we believe a constructive approach might leverage off the credibility of quasi-experimental estimates, and use them as inputs in an extrapolative exercise that will necessarily involve imposing more structure on the problem.

4. RESEARCH DESIGNS DOMINATED BY SELF-SELECTION

In this section, we briefly consider a group of research designs where institutional knowledge of the treatment assignment process typically does not provide most of the information needed to draw causal inferences. In Section 3, we discussed how some

[38] Note that $\int [(p_1^*(w(u), u) - p_0^*(w(u), u)) \frac{f_{V|U=u}(0)}{f_V(0)}] \mathrm{d}F_U(u) = \int p_1^*(w(u), u) \frac{f_{V|U=u}(0)}{f_V(0)} \mathrm{d}F_U(u) - \lim_{v\uparrow 0} \int p_0^*(w(u), u) \frac{f_{V|U=u}(v)}{f_V(v)} \mathrm{d}F_U(u) = \int p_1^*(w(u), u) \mathrm{d}F_{U|V=0}(u) - \lim_{v\uparrow 0} \int p_0^*(w(u), u) \mathrm{d}F_{U|V=v}(u) = E[D|V=0] - \lim_{v\uparrow 0} E[D|V=v]$.

aspects of the assignment process could be treated more as literal "descriptions" ("D"-conditions), rather than conjectures or assumptions. In each of the four research designs, those "D" conditions went a long way towards identification, and when other structural assumptions ("S"-conditions) were needed, the class of models consistent with those assumptions, while strictly smaller because of the restrictions, arguably remained very broad.

With the common program evaluation approaches considered in this section, we shall see that the assignment process is not dominated by explicit institutional knowledge, and identification thus requires more conjectures/assumptions ("S"-conditions) to make causal inferences. We will argue that with these designs there will be more scope for alternative plausible economic models that would be strictly inconsistent with the conditions needed for identification. Of the three approaches we consider, the "difference-in-difference" approach appears to have the best potential for testing the key "S"-conditions needed for identification.

For the reasons above, we suggest that these designs will tend to deliver causal inferences with lower internal validity, in comparison to the designs described in Section 3. But even if one agrees with the three criteria we have put forth in Section 2.2.1 to assess internal validity, and even if one agrees with the conclusion that these designs deliver lower internal validity, the question of "how much lower" requires a subjective judgment, and such a question is ill-defined at any rate (what is a unit of "internal validity"?). Of the three criteria we discuss, the extent to which the conditions for identification can be treated as a hypothesis with testable implications seems to be the least subjective in nature.

In our discussion below, it is still true that a particular weighted average effect of "interest" from an ex ante evaluation problem may in general be quite different from the effects identified by these research designs. In this sense, these approaches suffer from similar "external validity" concerns as discussed in Section 3. We will therefore focus our discussion on the ex post evaluation goal, and do not have separate sections on ways to extrapolate from the results of an ex post evaluation to forecast the effect of interest in an ex ante evaluation.

4.1. Using longitudinal data: "difference-in-difference"

We now consider the case where one has longitudinal data on program participants and non-participants. Suppose we are interested in the effectiveness of a job training program in raising earnings. Y is now earnings and D is participation in the job training program. A commonly used approach in program evaluation is the "difference-in-difference" design, which has been discussed as a methodology and utilized in program evaluation research in some form or another countless times. Our only purpose here is to discuss how the design fits into the general framework we have used in this chapter, and to be explicit about the restrictions in a way that facilitates comparison with the designs in Section 3.

First, let us simplify the problem by considering the situation where the program was made available at only one point in time τ. This allows us to define $D = 1$ as those who were treated at time τ, and $D = 0$ as those who did not take up the program at that time.

- D7 (Program exposure at one point in time): Individuals will have $D = 0$ for all $t < \tau$; for $t \geq \tau$, the non-treated will continue to have $D = 0$ while the treated will have $D = 1$.

W will continue to denote all the factors that could potentially affect Y. Additionally, we imagine that this vector of variables could be partitioned into sub-vectors W_t, $t = 1, \ldots, \tau$. where the subscript denotes the value of the variables at time t.

Furthermore, we explicitly include time in the outcome equation as

$$Y_t = y(D, W, U, t).$$

A "difference-in-difference" approach begins by putting some structure on Y:

- S15 (Additive Separability): $y(D, W, U, t) = g(D, W, U) + \alpha(W, U) + h(W, t)$.

This highly restrictive structure (although it does not rule out heterogeneous treatment effects) is the standard "individual fixed effects" specification for the outcome, where $\alpha(W, U)$ captures the permanent component of the outcome.

Perhaps the most important thing to keep in mind is that D7 and S15 is *not* generally sufficient for the "difference-in-difference" approach to identify the treatment effects. This is because the two differences in question are

$$\begin{aligned} E[Y_\tau - Y_{\tau-k}|D = 1] &= E[g(1, W, U) - g(0, W, U)|D = 1] \\ &\quad + \int h(W, \tau) - h(W, \tau - k)\mathrm{d}F_{U|D=1}(u) \\ E[Y_\tau - Y_{\tau-k}|D = 0] &= \int h(W, \tau) - h(W, \tau - k)\mathrm{d}F_{U|D=0}(u). \end{aligned}$$

The term $E[g(1, W, U) - g(0, W, U)|D = 1]$ is equal to $E[y(1, W, U, \tau) - y(0, W, U, \tau)|D = 1]$, the treatment on the treated (TOT) parameter. When the second equation is subtracted from the first, the terms with h do not cancel without further restrictions.

One approach to this problem is to further assume that

- S16 (Influence of "Other factors" Fixed): $h(W, t) - h(W, t - k) = \gamma(t, t - k)$, for any k.

This certainly would ensure that the D-in-D estimand identifies the TOT. It is, however, restrictive: even if $h(W, t) = h(W_t)$—so that only contemporaneous factors are relevant—then as long as there were some factors in W that changed over time, this would be violated. Note that in this case, it is irrelevant how similar or different the distribution

of unobservable types are between the treated and non-treated individuals ($F_{U|D=1}(u)$ vs. $F_{U|D=0}(u)$).[39]

4.1.1. Assessment

In terms of our three criterion for assessing this approach, how does the D-in-D fare? It should be very clear from the above derivation that the model of both the outcome and treatment assignment is a far cry from a literal description of the data generating process, except for D7, which describes the timing of the program and structure of the data. As for the second criterion, it is not difficult to imagine writing down economic models that would violate the restrictions. Indeed, much of the early program evaluation literature (Heckman and Robb, 1985; Ashenfelter, 1978; Ashenfelter and Card, 1985) discussed different scenarios under which S15 and S16 would be violated, and how to nevertheless identify program effects.

On the other hand, there is one positive feature of this approach—and it is driven by S16, which is precisely the assumption that allowed identification of the program effect—is that there are strong testable predictions of the design, namely

$$\begin{aligned} &E[Y_\tau - Y_{\tau-k}|D=1] - E[Y_\tau - Y_{\tau-k}|D=0] \\ &= E[g(1,W,U) - g(0,W,U)|D=1] \end{aligned}$$

for all $k > 0$. That is, the choice of the base year in constructing the DD should be irrelevant. Put differently, it means that

$$E[Y_{\tau-k} - Y_{\tau-k-j}|D=1] - E[Y_{\tau-k} - Y_{\tau-k-j}|D=0] = 0$$

for $j, k > 0$: the DD estimand during the "pre-program" period should equal zero. As is well-known from the literature, there are as many testable restrictions as there are pre-program periods, and it is intuitive that the more evidence that these restrictions are not rejected, the greater confidence we might have in the causal inferences that are made from the D-in-D.

Overall, while it is clear that the assumptions given by S15 and S16 require a great deal of speculation—and arguably a greater suspension of disbelief, relative to the conditions outlined in Section 3—at least there is empirical evidence (the pre-program data) that can be used to assess the plausibility of the key identifying assumption, S16.

4.2. Selection on unobservables and instrumental variables

In this section, we briefly discuss the identification of program effects when we have much less information about treatment assignment, relative to the designs in Section 3.

[39] S16 has the implication that there will be no variance in $Y_\tau - Y_{\tau-k}$ for the untreated group, which will in practice almost never be the case. The variance in changes could be accommodated by an independent, additive error term in the outcome equation.

We will focus on the use of instrumental variable approaches, but it will be clear that the key points we discuss will equally apply to a control function approach.

The instrumental variable approach is typically described as finding a variable Z that impacts Y only through its effect on treatment status D. Returning to our job search assistance program example, let us take Z to be the binary variable of whether the individual's sibling participated in the program. The "story" behind this instrument would be that the sibling's participation might be correlated with the individual's participation—perhaps because the sibling would be an influential source of information about the program—but that there is no reason why a sibling's participation in the program would directly impact the individual's re-employment probabilities.

Even if one completely accepts this "story"—and there are undoubtedly reasons to question its plausibility—this is not sufficient to identify the treatment effect via this instrument. To see this, we use the framework in Eqs (1)–(3), and adopt S8 (Excludability) and S9 (Probabilistic Monotonicity).

S8 is the formal way to say that for any individual type U, Z (the sibling's program participation) has no direct impact on the outcome. This exclusion restriction might come from a particular behavioral model. Furthermore, S9 simply formalizes the notion that for any given type U, the probability of receiving treatment is higher when the sibling participated in the program. Alternatively, it says that for each type U, there are more individuals who are induced to receive treatment because of their sibling's participation, than those who would be discouraged from doing so.

The problem is that, in general, it is easy to imagine that there is heterogeneity in the latent propensity for the individual to have a sibling participate in the program: $P_Z^* = \Pr[Z = 1|U]$ has a non-degenerate distribution. If such variability in P_Z^* exists in the population, this immediately implies that $F_{U|Z=1}(u)$ will in general be different from $F_{U|Z=0}(u)$. The IV (Wald) estimand will in general not identify any average effect, LATE or otherwise.

Typically researchers immediately recognize that their instrument Z is not "as good as randomly assigned" as in D3, and so instead appeal to a "weaker" condition that

- S17 (Conditional on W, Z "as good as randomly assigned"): $\Pr[Z = 1|U = u] = p_{z1}(w(u))$, a function of w.

This is a restriction on the heterogeneity in P_Z^*. S17 says that types with the same W have identical propensities P_Z^*.

Of course, the notion that the analyst knows and could measure all the factors in W, is a conjecture in itself:

- S18 (Sufficient variables for P_Z^*): Let $X = x(U)$ be the observable (to the researcher) elements of W, and assume $p_{z1}(w(u)) = p_{z1x}(x(u))$ for all u.

S18 simply says that the researcher happens to observe all the variables that determine the propensity P_Z^*.

It should be clear that with S17, S18, S8, and S9, one can condition the analysis on a particular value $X = x$, and apply the results from the randomized experiment with imperfect compliance (Section 3.2).

Note that while we have focused on the binary instrument case, it should also be clear that this argument will apply to the case when Z is continuously distributed. It is not sufficient for Z to simply be excluded from the outcome equation, Z must be assumed to be (conditionally) independent of U, and indeed this is the standard assumption in the evaluation literature[40]

4.2.1. Assessment

It is clear that in this situation, as D3 is replaced with S17 and S18, there is now *no* element of the statistical model that can be considered a literal description of the treatment assignment process. The model is entirely driven by assumptions about economic behavior, rather than a description of the data generating process that is derived from our institutional knowledge about how treatment was assigned.

S17 makes it clear that causal inferences will be dependent on having the correct set of variables X. Without the complete set of X, there will be variability in P_Z^* conditional on the covariates, which will mean that the distribution of types U will not be the same in the $Z = 1$ and $Z = 0$ groups. In general, different theories about which X's satisfy S18 will lead to a different causal inference. Recall that no similar specification of the relevant X's was necessary in the case of the randomized experiment with imperfect compliance, considered in Section 3.2.

Finally, there seems to be very little scope for testing the validity of this design. If the argument is that the instrument is independent of U only conditional on all the Xs observed by the researcher, then all the data will have been "used up" to identify the causal parameter.

To make the design somewhat testable, the researcher could assume that only a smaller subset of variables in X are needed to characterize the heterogeneity in P_Z^*. In that case, one could imagine conditioning on that smaller subset, and examining whether the distribution of the remaining X variables are balanced between the $Z = 1$ and $Z = 0$, as suggested in Section 3.2 (with the appropriate caveats and qualifications discussed there). But in practice, when evidence of imbalance is found, the temptation is to simply include those variables as conditioning variables to achieve identification, which then eliminates the potential for testing.

What this shows is the benefit to credibility that one obtains from explicit knowledge about the assignment process whereby D3 is a literal description of what is known about the process. It disciplines the analysis so that any observed X variables can be used to treat D3 as a hypothesis to be tested.

[40] In the more general discussion in Heckman and Vytlacil (2005), for example, Z is assumed, at a minimum, to be independent of U_1 and U_0 (or of potential outcomes) given a set of observed "conditioning variables."

4.3. Selection on observables and matching

Absent detailed institutional knowledge of the selection process, i.e., the propensity score equation, a common approach to the evaluation problem is to "control" for covariates, either in a multiple regression framework, or more flexibly through multivariate matching or propensity score techniques.[41] Each of these approaches essentially assumes that conditional on some observed covariates X, treatment status is essentially "as good as randomly assigned".

In terms of the model given in Eqs (1)–(3), one can think of the selection on observables approach as amounting from two important assumptions. First we have

- S19 (Conditional on W, D is "as good as randomly assigned"): $P^* = p^*(w(U))$, a function of w.

Here, the unobservable type U does not enter the latent propensity, whereas it does in (2).

It is important to reiterate that the function $\Pr[D = 1|W = w]$, the so-called "propensity score"—which can be obtained as long as one can observe W—is *not,* in general, the same thing as the latent propensity P^* for a given $W = w$. That is, even though it will always be true that $\Pr[D = 1|W = w] = E[P^*|W = w]$, there will be heterogeneity in P^* for a given $W = w$, unless one imposes the condition S19. And it is precisely this heterogeneity, and its correlation with the outcome Y, that is the central problem of making causal inferences, as discussed in Section 2.2.

In addition, if the researcher presumes that there are some factors that are unobservable in W, then one must further assume that

- S20 (S19 + Sufficient variables for P^*): Let $X = x(U)$ be the observable (to the researcher) elements of W, and assume $p^*(w(u)) = p^*_x(x(u))$ for all u.

So S20 goes further to say that not only are the Xs sufficient to characterize the underlying propensity, all the unobservable elements of W are irrelevant in determining the underlying propensity P^*.

S20 has the same implications as condition D2 discussed in Section 3.1.2: the difference $E[Y|D = 1, X = x] - E[Y|D = 0, X = x]$ identifies the (conditional) average treatment effect $E[\Delta(W, U)|X = x]$. The key difference is that in Section 3.1.2, D2 was a literal description of a particular assignment process (random assignment with probabilities of assignment being a function of X).[42] Here, S20 is a restriction on the framework defined by Eqs (1)–(3).

To see how important it is not to have variability in P^* conditional on X, consider the "conditional" version of Eq. (5)

[41] For a discussion of several of these approaches, see Busso et al. (2008, 2009).

[42] Indeed, for a "half century" the basic framework was "entirely tied to randomization based evaluations" and was "not perceived as being relevant for defining causal effects in observational studies." Rubin (1990).

$$
\begin{aligned}
&E[Y|D=1, X=x] - E[Y|D=0, X=x] \\
&\quad = \int E[y(1, w(U), U)|P^* = p^*, X = x] f_{P^*|D=1,X=x}(p^*)\mathrm{d}p^* \\
&\qquad - \int E[y(0, w(U), U)|P^* = p^*, X = x] f_{P^*|D=0,X=x}(p^*)\mathrm{d}p^*.
\end{aligned}
$$

A non-degenerate density $f_{P^*|X=x}(p^*)$ will automatically lead to $f_{P^*|D=1,X=x}(p^*)\mathrm{d}p^* \neq f_{P^*|D=0,X=x}(p^*)\mathrm{d}p^*$, which would prevent the two terms from being combined.[43]

4.3.1. Assessment: included variable bias

In most observational studies, analysts will rarely claim that they have a model of behavior or institutions that dictate that the assignment mechanism *must* be modeled as S20. More often, S20 is invoked because there is an explicit recognition that there is non-random selection into treatment, so that D is certainly not unconditionally randomly assigned. S20 is offered as a "weaker" alternative.

Perhaps the most unattractive feature of this design is that, even if one believes that S20 does hold, typically there is not much in the way of guidance as to what Xs to include, as has long been recognized (Heckman et al., 1998). There usually is a multitude of different plausible specifications, and no disciplined way to choose among those specifications.

It is therefore tempting to believe that if we compare treatment and control individuals who look more and more similar on observable dimensions X, then—even if the resulting bias is non-zero—at the least, the bias in the estimate will decrease. Indeed, there is a "folklore" in the literature which suggests that "overfitting" is either beneficial or at worst harmless. Rubin and Thomas (1996) suggest including variables in the propensity score unless there is a consensus that they do not belong. Millimet and Tchernis (2009) go further and argue that *overfitting* the propensity score equation is possibly beneficial and at worst harmless.

It is instructive to consider a few examples to illuminate why this is in general not true, and how adding X's can lead to "included variable bias". To gain some intuition, first consider the simple linear model

$$Y = \beta_0 + D\beta_1 + \epsilon$$

where β_1 is the coefficient of interest and $\mathrm{COV}(D, \epsilon) \neq 0$. The probability limit of the OLS regression coefficient on D is

$$\beta_{OLS} = \beta_1 + \frac{\mathrm{COV}(D, \epsilon)}{\mathrm{VAR}(D)}.$$

[43] Recall that $f_{P^*|D=1,X=x}(p^*)\mathrm{d}p^* = \frac{p^*}{\Pr[D=1|X=x]} f_{P^*|X=x}(p^*)\mathrm{d}p^*$ and $f_{P^*|D=0,X=x}(p^*)\mathrm{d}p^* = \frac{1-p^*}{\Pr[D=0|X=x]} f_{P^*|X=x}(p^*)\mathrm{d}p^*$, which will be unequal with $f_{P^*}(p^*)$ non-degenerate.

Now suppose there is a "control variable" X. Suppose X actually has covariance $\text{COV}(X, \epsilon) = 0$, so it is an "irrelevant" variable, but it can explain some variation in D. When X is included, the least squares coefficient on D will be

$$\beta_{OLS,X} = \frac{\text{COV}(Y, D - \hat{D})}{\text{VAR}(D - \hat{D})} = \beta_1 + \frac{\text{COV}(D, \epsilon)}{\text{VAR}(D - \hat{D})}$$

where $\hat{D}$ is the predicted value from a population regression of D on X. This expression shows that the magnitude of the bias in the least squares estimand that includes X will be strictly larger, with the denominator in the bias term decreasing. What is happening is that the extra variable X is doing nothing to reduce bias, while absorbing some of the variation in D.

To gain further intuition on the potential harm in "matching on Xs" in the treatment evaluation problem, consider the following system of equations

$$\begin{aligned} Y &= \beta_0 + D\beta_1 + X\beta_2 + U \\ D &= 1[\delta_0 + X\delta_1 + V \geq 0] \end{aligned}$$

where X, the "control" variable, is in this case a binary variable. (U, V) is assumed to be independent of X. This is a simplified linear and parametric version of (1), (2), and (3).[44]

The bias of the simple difference in means—without accounting for X—can be shown to be

$$\begin{aligned} \text{BIAS}_{\text{DIF}} &\equiv E[Y|D = 1] - E[Y|D = 0] - \beta_1 \\ &= \beta_2(\Pr[X = 1|D = 1] - \Pr[X = 1|D = 0]) \\ &\quad + \{E[U|D = 1] - E[U|D = 0]\} \end{aligned}$$

whereas the bias in the matching estimand for the TOT is

$$\begin{aligned} &\text{BIAS}_{\text{MATCH}} \\ &\equiv \left[\sum_x \Pr[X = x|D = 1]\,(E[Y|D = 1, X = x] - E[Y|D = 0, X = x])\right] - \beta_1 \\ &= \left[\sum_x \Pr[X = x|D = 1]\,(E[U|D = 1, X = x] - E[U|D = 0, X = x])\right] \\ &= E[U|D = 1] - (E[U|D = 0, X = 0]\Pr[X = 0|D = 1] \\ &\quad + E[U|D = 0, X = 1]\Pr[X = 1|D = 1]). \end{aligned}$$

[44] Here, $P^* = \Pr[V \geq -\delta_0 - X\delta_1|X, U]$.

In this very simple example, a comparison between BIAS_{DIF} and $\text{BIAS}_{\text{MATCH}}$ reveals two sources of differences. First, there is a standard "omitted variable bias" that stems from the first term in BIAS_{DIF}. This bias does not exist in the matching estimand.

But there is another component in BIAS_{DIF}, the term in curly braces—call it the "selectivity bias" term. This term *will always be smaller in magnitude than* $|\text{BIAS}_{\text{MATCH}}|$. That is, "controlling for" X *can only increase* the magnitude of the selectivity bias term. To see this, note that the difference between the term in curly braces in BIAS_{DIF} and $\text{BIAS}_{\text{MATCH}}$ is the difference between

$$\begin{aligned} E[U|D=0] = {} & E[U|D=0, X=0]\Pr[X=0|D=0] \\ & + E[U|D=0, X=1]\Pr[X=1|D=0] \end{aligned} \tag{15}$$

and

$$E[U|D=0, X=0]\Pr[X=0|D=1] + E[U|D=0, X=1]\Pr[X=1|D=1] \tag{16}$$

in $\text{BIAS}_{\text{MATCH}}$. Each of these expressions is a weighted average of $E[U|D=0, X=x]$.

Consider the case of positive selectivity, so $E[U|V=v]$ is increasing in v, so that the selectivity term (curly braces) in BIAS_{DIF} is positive, and suppose that $\Pr[D=1|X=1] > \Pr[D=1|X=0]$ (i.e. $\delta_1 > 0$).[45] This means that $E[U|D=0, X=1] = E[U|V < -\delta_0 - \delta_1] < E[U|V < -\delta_0] = E[U|D=0, X=0]$. That is, among the non-treated group, those with $X=1$ are *more* negatively selected than those with $X=0$.

Comparing (15) and (16), it is clear that $\text{BIAS}_{\text{MATCH}}$ automatically places relatively more weight on $E[U|D=0, X=1]$, since $\Pr[X=1|D=1] > \Pr[X=1|D=0]$, and hence $\text{BIAS}_{\text{MATCH}}$ will exceed the selectivity term (curly braces) in BIAS_{DIF}.[46] Intuitively, as X can "explain" more and more of the variation in D, the exceptions (those with $X=1$, but $D=0$) must have unobservable factors that are even more extreme in order to be exceptional. And it is precisely those exceptional individuals that are implicitly given relatively more weight when we "control" for X.

So in the presence of nontrivial selection on unobservables, a matching on observables approach will generally *exacerbate* the selectivity bias. Overall, this implies that a reduction in bias will require the possible "benefits"—elimination of the omitted variable bias driven by β_2—to outweigh the cost of exacerbating the selectivity bias.

There is another distinct reason why the magnitude of $\text{BIAS}_{\text{MATCH}}$ may be larger than that of BIAS_{DIF}. The problem is that β_2 is unknown, and the sign and magnitude

[45] Parallel arguments hold when $E[U|V=v]$ is decreasing in v and/or when $\delta_1 < 0$.

[46] By Bayes' rule, $\Pr[X=1|D=1] = \frac{\Pr[D=1|X=1]}{\Pr[D=1]}\Pr[X=1] > \frac{1-\Pr[D=1|X=1]}{1-\Pr[D=1]}\Pr[X=1] = \Pr[X=1|D=0]$.

need not be tied to the fact that U correlates with V. In the above example, even if there is positive selectivity on unobservables, β_2 may well be negative, and therefore, BIAS_{DIF} could be zero (or very small). So even if matching on X had a small effect on the selectivity bias component, the elimination of the omitted variable bias term will cause $\text{BIAS}_{\text{MATCH}} > \text{BIAS}_{\text{DIF}}$. That is, if the two sources of biases were offsetting each other in the simple difference, eliminating one of the problems via matching make the overall bias increase.

Overall, we conclude that as soon as the researcher admits departures from S20, there is a rather weak case to be made for "matching" on observables being an improvement. Indeed, there is a compelling argument that including more X's will increase bias, and that the "cure may be worse than the disease".

Finally, in terms of the third criterion we have been considering, the matching approach seems to have no testable implications whatsoever. One possibility is to specify a particular subset X' of the Xs that are available to the researcher, and make the argument that it is specifically those variables that determine treatment in S20. The remainder of the observed variables could be used to test the implication that the distribution of types U is the same between the treated and non-treated populations, conditional on X'. The problem, of course, is that if some differences were found in those Xs not in X', there would again be the temptation to simply include those variables in the subset X'. Overall, not only do we believe this design to have a poor theoretical justification in most contexts (outside of actual stratified randomized experiments), but there seems to be nothing in the design to discipline which Xs to include in the analysis, and as we have shown above, there is a great risk to simply adding as many Xs to the analysis as possible.

4.3.2. *Propensity score, matching, re-weighting: methods for descriptive, non-causal inference*

Although we have argued that the matching approach is not compelling as a research design for causal inference, it can nevertheless be a useful tool for descriptive purposes. Returning to our hypothetical job search assistance program, suppose that the program is voluntary, and that none of the data generating processes described in Section 4 apply. We might observe the difference

$$E[Y|D=1] - E[Y|D=0]$$

but also notice that the distribution of particular Xs (education, age, gender, previous employment history) are also different: $F_{X|D=1}(x) \neq F_{X|D=0}(x)$. One could ask the descriptive question, "mechanically, how much of the difference could be *exclusively* explained by differences in the distribution of X?" We emphasize the word "mechanically"; if we observe that Y varies systematically by different values of X for the treated population, and if we further know that the distribution of X is different in

the non-treated population, then even if the program were entirely irrelevant, we would nevertheless generally *expect* to see a difference between $E[Y|D=1]$ and $E[Y|D=0]$.

Suppose we computed

$$\widehat{E[Y|D=0]} \equiv \int E[Y|D=0, X=x]\mathrm{d}F_{X|D=1}(x).$$

Then the difference

$$E[Y|D=1] - \widehat{E[Y|D=0]} \tag{17}$$

would tell us how relevant D is in predicting Y, once adjusting for the observables X. If one adopted S20 then this could be interpreted as an average treatment effect for the treated. But more generally, this adjusted difference could be viewed as a descriptive, summary statistic, in the same way multiple regressions could similarly provide descriptive information about the association between Y and D after partialling out X.

We only briefly review some of the methods used to estimate the quantity (17), since this empirical exercise is one of the goals of more general decomposition methods, which is the focus of the chapter by Firpo et al. (2011). We refer the reader to that chapter for further details.

Imputation: Blinder/Oaxaca

One way of obtaining (17), is to take each individual in the treated sample, and "impute" the missing quantity, the average Y given the individual's characteristics X. This is motivated by the fact that

$$\int\int [y - E[Y|D=0, X=x]] f_{X,Y|D=1}(x,y)\mathrm{d}x\mathrm{d}y \tag{18}$$

is identical to the quantity (17).

The sample analogue is given by

$$\frac{1}{N_1}\sum_{i:D_i=1}[Y_i - \widehat{Y}_i]$$

where N_1 is the number of observations in the treated sample, and $\hat{Y}_i$ is the predicted value of regressing Y on X for the non-treated sample. This is immediately recognizable as a standard Blinder/Oaxaca exercise.

Matching

One development in the recent labor economics literature is an increased use of matching estimators, estimators based on the propensity score and semi-parametric estimators

which eschew parametric specification of the outcome functions. The concern is that the regression used to predict $\hat{Y}_i$ may be a bad approximation of the true conditional expectation.

The first approach is to simply use the sample mean of Y_i for all individuals in the non-treated sample that have exactly the same value for X as the individual i. Sometimes it will be possible to do this for every individual (e.g. when X is discrete and for each value of X there are treated and non-treated observations).

In other cases, X is so multi-dimensional that for each value of X there are very few observations with many values only having treated or non-treated observations. Alternatively, X could have continuously distributed elements, in which case exact matching is impossible. In this case, one approach is to compute non-parametric estimates of $\hat{Y}_i$ using kernel regression or local polynomial regression (Hahn, 1998; Hirano et al., 2003). A version of matching takes the data point in the control sample that is "closest" to the individual i in terms of the characteristics X, and assigns $\hat{Y}_i$ to be the value of Y for that "nearest match".

Propensity score matching

A variant of the above matching approach is to "match" on the Propensity score, rather than on the observed X, and it is motivated by the fact that (17) is also equivalent to

$$\int \int [y - E[Y|D = 0, PS(x) = p]] f_{PS,Y|D=1}(p, y) \mathrm{d}p \mathrm{d}y$$

where

$$PS(x) = \Pr[D = 1|X = x]$$

is the well-known "propensity score" of Rosenbaum and Rubin (1983). We emphasize once again that PS is *not* the same thing as P^*, the latent propensity to be treated. Indeed it is the fact that there may be variability in P^* conditional on PS, which threatens the validity of the "selection on observables" approach to causal inference.

Re-weighting

An alternative approach is to "re-weight" the control sample so that the re-weighted distribution of X matches that in the treated population. It is motivated by the fact that (18) is also equivalent to

$$E[Y|D = 1] - \int \left(\int y f_{Y|X=x,D=0}(y) \mathrm{d}y \right) f_{X|D=1}(x) \mathrm{d}x$$

which is equal to

$$E[Y|D=1] - \int\int y f_{Y|X=x,D=0}(y) \frac{f_{X|D=1}(x)}{f_{X,Y|D=0}(x,y)} f_{X,Y|D=0}(x,y)\mathrm{d}x\mathrm{d}y.$$

Using the fact that $f_{Y|X=x,D=0}(y) = \frac{f_{X,Y|D=0}(x,y)}{f_{X|D=0}(x)}$, this becomes

$$E[Y|D=1] - \int\int y \left(\frac{f_{X|D=1}(x)}{f_{X|D=0}(x)}\right) f_{X,Y|D=0}(x,y)\mathrm{d}x\mathrm{d}y.$$

The second term is simply a weighted average of Y for the non-treated observations using $\left(\frac{f_{X|D=1}(x)}{f_{X|D=0}(x)}\right)$ as a weight. It is clear that this average will up-weight those individuals with $X = x$, when relatively "more" individuals with that value are among the treated than among the non-treated; when there are disproportionately fewer individuals with $X = x$, the weighted average will down-weight the observation.

By Bayes' rule, this weight is also equal to

$$\frac{\Pr[D=1|X=x]}{1-\Pr[D=1|X=x]} \frac{1-\Pr[D=1]}{\Pr[D=1]} = \frac{PS(x)}{1-PS(x)} \frac{1-\Pr[D=1]}{\Pr[D=1]}$$

(DiNardo et al., 1996; Firpo, 2007).

Thus, in practice, a re-weighting approach will involve computing the sample analogue

$$\frac{1}{N_1} \sum_{i:D_i=1} Y_i - \frac{1}{N_0} \sum_{i:D_i=0} \frac{\hat{PS}(X_i)}{1-\hat{PS}(X_i)} \frac{N_0}{N_1} Y_i$$

where $\hat{PS}(x)$ is the estimated propensity score function for an individual i with $X_i = x$.

A useful aspect of viewing the adjustment as a re-weighting problem is that one is not limited to examining only conditional expectations of Y: one can re-weight the data and examine other aspects of the distribution, such as quantiles, variances, etc. by computing the desired statistic with the appropriate weight. See (DiNardo et al., 1996; DiNardo and Lemieux, 1997; Biewen, 1999; Firpo, 2007) for discussion and applications.

5. PROGRAM EVALUATION: LESSONS AND CHALLENGES

This chapter provides a systematic assessment of a selection of commonly employed program evaluation approaches. We adopt a perspective that allows us to consider how the Regression Discontinuity Design—an approach that has seen a marked increase in use over the past decade—relates to other well-known research designs. In our discussion,

we find it helpful to make two distinctions. One is between the descriptive goals of an ex post evaluation, and the predictive goals of an ex ante evaluation. And the other is between two kinds of statistical conditions needed to make causal inference—(1) descriptions of our institutional knowledge of the program assignment process, and (2) structural assumptions—some that have testable restrictions, and others that will not—that do not come from our institutional knowledge, but rather stem from conjectures and theories about individual behavior; such structural assumptions necessarily restrict the set of models of behavior within which we can consider the causal inference to be valid.

In our discussion, we provide three concrete illustrations of how the goals of ex post and ex ante evaluations are quite complementary. In the case of the randomized experiment with perfect compliance, highly credible estimates can be obtained for program effects for those who selected to be a participant in the study. Through the imposition of a number of structural assumptions about the nature of the economy, one can draw a precise link between the experimentally obtained treatment effect and a particular policy parameter of interest—the aggregate impact of a wide-spread "scaling" up of the program. In the case of the randomized experiment with imperfect compliance, one can make highly credible inferences about program effects, even if the obtained treatment effect is a weighted average. But with an additional functional form assumption (that is by no means unusual in the applied literature), one can extrapolate from a Local Average Treatment Effect to the Average Treatment Effect, which might be the "parameter of interest" in an ex ante evaluation. Finally, in the case of the RD design, one can obtain highly credible estimates of a weighted average treatment effect, which in turn can be viewed as an ingredient to an extrapolation for the Treatment on the Treated parameter.

Our other observation is that "D"-conditions and "S"-conditions are also quite complementary. On the one hand, for the designs we examine above, "D"-conditions are generally necessary (even if not sufficient) to isolate the component of variation in program status that is "as good as randomly assigned". When they are not sufficient, "S"-conditions are needed to fill in the missing pieces of the assignment process. Furthermore, in our three illustrations, only with "S"-conditions can any progress be made to learn about other parameters of interest defined by an ex ante evaluation problem. Thus, our three examples are not meant to be definitive, but rather illustrative of how research designs dominated by "D"-conditions could supply the core ingredients to ex ante evaluations that are defined by "S"-conditions. In our view, this combination seems promising.

More importantly, what is the alternative? There is no a priori reason to expect that the variation that we may be able to isolate as "effectively randomized", to be precisely the variation required to identify a particular policy proposal of interest, particularly since what is "of interest" is subjective, and researcher-dependent.[47] That is, in virtually any

[47] Heckman and Vytlacil (2005) make the point that if the propensity score has limited support (e.g. including discrete support), marginal treatment effects cannot be identified in certain areas, and certain policy parameters of interest are also not identified.

context—experimental or non-experimental—the effects we *can* obtain will not exactly match what we *want*. The alternative to being precise about the sub-population for whom the effects are identified is to be imprecise about it. And for conducting an ex ante evaluation, the alternative to using an extrapolation where the leading term is a highly credible experimental/quasi-experimental estimate is to abandon that estimate in favor of an extrapolation in which the leading term is an estimate with questionable or doubtful internal validity. Similarly, even if the assumptions needed for extrapolation involve structural assumptions that require an uncomfortable suspension of disbelief, the alternative to being precise in specifying those assumptions, is to be imprecise about it and make unjustified generalizations, or to abandon the ex ante evaluation question entirely.

We conclude with some speculation on what could be fruitful directions for further developing strategies for the ex post evaluation problem. One observation from our discussion is that both the Sharp RD and Fuzzy RD are representations of the general self-selection problem, where agents can take actions to influence their eligibility or participation in a program. What allows identification—and indeed the potential to generate randomization from a non-experimental setting—is our *knowledge* of the threshold, and the *observability* of the "latent variable" that determines the selection. Turning that on its head, we could view *all* selection problems with a latent index structure as inherently Regression Discontinuity designs, but ones for which we do *not* perfectly observe the latent selection variable (or the cutoff). But what if we have partial institutional knowledge on the assignment process? That is, even if we don't measure V (from Sections 3.3 and 3.4), what if we observe a reasonable proxy for V? Can that information be used?

On a related point, our presentation of various research designs has a "knife-edge" quality. The designs in Section 3 are such that P^* or p_{z1} have point-mass distributions, or we required *every* individual to have a continuous density for V. When those conditions held, we argued that the effects would be highly credible, with strong, testable implications. But in Section 2.2.1, we argued that when we do *not* have specific knowledge about the assignment process, the designs will tend to yield more questionable inferences, because there will be an increase in plausible alternative specifications often if with very little to guide us as to the "preferred" specification. Does a middle-ground exist? Are there situations where our knowledge of the assignment process tells us that P^* or p_{z1}, while not distributed as a mass-point, has *small* variance? Might there be ways to adjust for these "minor" departures from "as good as randomized"?

Finally, another lesson from the RD design is how much is gained from actually knowing something about the treatment assignment process. It is intuitive that when one actually knows the rule that partially determines program status, and one observes the selection rule variable, that should help matters. And it is intuitive that if program assignment is a complete "black box"—as is often the case when researchers invoke a "selection on observables"/matching approach—we will be much less confident about

those program effects; one ought to be a bit skeptical about strong claims to the contrary. Since most programs are at least partially governed by some eligibility rules, the question is whether there are some other aspects of those rules—that go beyond discontinuities or actual random assignment—from which we can tease out credible inferences on the programs' causal impacts.

REFERENCES

Abadie, Alberto, Angrist, Joshua D., Imbens, Guido, 2002. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. Econometrica 70 (1), 91–117.

Abbring, Jaap H., Heckman, James J., 2007. Econometric evaluation of social programs, Part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. In: Heckman, J.J., Leamer, E.E. (Eds.), Handbook of Econometrics. In: Handbook of Econometrics, vol. 6. Elsevier (Chapter 72).

Angrist, Joshua D., 1990. Lifetime earnings and the Vietnam Era draft lottery: Evidence from social security administrative records. American Economic Review 80 (3), 313–336.

Angrist, Joshua D., 2004. Treatment effect heterogeneity in theory and practice. Economic Journal 114 (494), C52–C83.

Angrist, Joshua D., Krueger, Alan B., 1999. Empirical strategies in labor economics. In: Ashenfelter, Orley, Card, David (Eds.), Handbook of Labor Economics. In: Handbooks in Economics, vol. 3-A. Elsevier Science, New York, pp. 1277–1366 (Chapter 23).

Angrist, Joshua D., Pischke, Jörn-Steffen, The credibility revolution in empirical economics: how better research design is taking the con out of econometrics. NBER Working Papers 15794, National Bureau of Economic Research, Inc. March 2010.

Angrist, Joshua D., Lavy, Victor, 1999. Using Maimonides' rule to estimate the effect of class size on scholastic achievement. Quarterly Journal of Economics 114 (2), 533–575.

Angrist, Joshua D., Evans, William N., 1998. Children and their parents' labor supply: evidence from exogenous variation in family size. American Economic Review 88 (3), 450–477.

Angrist, Joshua D., Imbens, Guido W., Rubin, Donald B., 1996. Identification of causal effects using instrumental variables. Journal of the American Statistical Association 91 (434), 444–455.

Angrist, Joshua, Bettinger, Eric, Kremer, Michael, 2006. Long-term educational consequences of secondary school vouchers: evidence from administrative records in Colombia. American Economic Review 96 (3), 847–862.

Ashenfelter, Orley, 1978. Estimating the effect of training programs on earnings. Review of Economics and Statistics 60 (1), 47–57.

Ashenfelter, Orley, Card, David, 1982. Time series representations of economic variables and alternative models of the labour market. Review of Economic Studies 49 (5), 761–781.

Ashenfelter, Orley, Card, David, 1985. Using the longitudinal structure of earnings to estimate the effect of training programs. Review of Economics and Statistics 67.

Ashenfelter, Orley, Plant, Mark W., 1990. Nonparametric estimates of the labor-supply effects of negative income tax programs. Journal of Labor Economics 8 (1), S396–S415.

Barnow, B., Cain, G., Goldberger, A., 1976. Issues in the analysis of selectivity bias. Evaluation Studies Review Annual 5, 43–59.

Biewen, Martin, Measuring the effects of socio–economic variables on the income distribution: an application to the east German transition process. Discussion Paper Series, Ruprecht–Karls–Universität, Heidelberg, Germany, March 1999.

Black, Sandra, 1999. Do better schools matter? parental valuation of elementary education. Quarterly Journal of Economics 114 (2), 577–599.

Busso, Matias, DiNardo, John, McCrary, Justin, 2008. Finite sample properties of semiparametric estimators of average treatment effects. Unpublished Working Paper, University of Michigan, Ann Arbor, MI. September 19.

Busso, Matias, DiNardo, John, McCrary, Justin, 2009. New evidence on the finite sample properties of propensity score matching and reweighting estimators. Working Paper 3998, Institute for the Study of Labor (IZA). February.

Campbell, Donald T., Cook, Thomas D., 1979. Quasi-Experimentation: Design and Analysis for Field Settings, first edition, Rand McNally College Publishing Company, Chicago.

Card, David, Krueger, Alan B., 1995. Myth and Measurement: The New Economics of the Minimum Wage. Princeton University Press, Princeton, NJ.

Card, David, Dobkin, Carlos, Maestas, Nicole, 2009a. Does Medicare save lives? Quarterly Journal of Economics 124 (2), 597–636.

Card, David, Dobkin, Carlos, Maestas, Nicole, 2009b. The impact of nearly universal insurance coverage on health care utilization: evidence from medicare. American Economic Review.

Cook, T.D., 2008. "Waiting for life to arrive": A history of the regression-discontinuity design in psychology, statistics and economics. Journal of Econometrics 142 (2), 636–654.

Cox, D.R., 1958. Planning of Experiments. Wiley, New York.

Deaton, Angus S., 2008. Instruments of Development: randomization in the tropics and the search for the elusive keys to development. Proceedings of the British Academy 162, 123–160. Keynes Lecture, British Academy.

DiNardo, John, Lee, David S., 2004. Economic impacts of new unionization on private sector employers: 1984–2001. Quarterly Journal of Economics 119 (4), 1383–1441.

DiNardo, John, Lemieux, Thomas, 1997. Diverging male wage inequality in the United States and Canada, 1981–1988: do institutions explain the difference? Industrial and Labor Relations Review.

DiNardo, John, Fortin, Nicole, Lemieux, Thomas, 1996. Labor market institutions and the distribution of wages, 1973–1993: a semi-parametric approach. Econometrica 64 (5), 1001–1045.

Fan, Jianqing, Gijbels, Irene, 1996. Local Polynomial Modelling and its Applications. Chapman and Hall, New York.

Fang, Hanming, Keane, Michael, Khwaja, Ahmed, Salm, Martin, Silverman, Daniel, 2007. Testing the mechanisms of structural models: The case of the mickey mantle effect. American Economic Review 97 (2), 53–59.

Fernández-Villaverde, Jesús, 2009. The Econometrics of DSGE Models. Working Paper 14677, National Bureau of Economic Research. January.

Field, Erica, 2007. Entitled to work: urban property rights and labor supply in Peru. The Quarterly Journal of Economics 122 (4), 1561–1602.

Firpo, Sergio, 2007. Efficient semiparametric estimation of quantile treatment effects. Econometrica 75 (1), 259–276.

Firpo, Sergio, Forin, Nicole, Lemieux, Thomas, 2011. Decomposition methods in economics. In: Ashenfelter, Orley, Card, David (Eds.), Handbook of Labor Economics, vol. 4A. North Holland, Amsterdam, pp. 1–102.

Fisher, Sir Ronald Aylmer, 1935. Design of Experiments. Oliver and Boyd, Edinburgh, London.

Fisher, Sir Ronald Aylmer, 1966. Design of Experiments, eighth ed., Oliver and Boyd, Edinburgh, London, First edition published in 1935.

Freedman, David A., 1983. A note on screening regression equations. The American Statistician 37 (2), 152–155.

Guttman, Robert, 1983. Job training partnership act: new help for the unemployed. Monthly Labor Review 3–10.

Haavelmo, Trygve, 1944. The probability approach in econometrics. Econometrica (12), iii–115.

Hacking, Ian, 1965. The Logic of Statistical Inference. Cambridge University Press, Cambridge.

Hahn, Jinyong, 1998. On the role of the propensity score in efficient semiparametric estim ation of average treatment effects. Econometrica 66 (2), 315–331.

Hahn, Jinyong, Todd, Petra, Van der Klaauw, Wilbert, 2001. Identification and estimation of treatment effects with a regression-discontinuity design. Econometrica 69 (1), 201–209.

Hearst, Norman, Newman, Tom B., Hulley, Stephen B., 1986. Delayed effects of the military draft on mortality: a randomized natural experiment. New England Journal of Medicine 314, 620–624.

Heckman, James J., 1974. Shadow prices, market wages, and labor supply. Econometrica 42 (4), 679–694.

Heckman, James J., 1976. The common structure of statistical models of truncation, sample selection, and limited dependent variables, and a simple estimator for such models. Annals of Economic and Social Measurement 5 (4), 475–492.
Heckman, James J., 1978. Dummy endogenous variables in a simultaneous equation system. Econometrica 46, 931–960.
Heckman, James J., 1991. Randomization and social policy evaluation. Working Paper 107. National Bureau of Economic Research. July.
Heckman, James J., 2000. Causal parameters and policy analysis in economics: a twentieth century retrospective. Quarterly Journal of Economics 115 (1), 45–97.
Heckman, James J., 2001. Micro data, heterogeneity, and the evaluation of public policy: nobel lecture. The Journal of Political Economy 109 (4), 673–748.
Heckman, James J., Honore, Bo E., 1990. The empirical content of the roy model. Econometrica 58 (5), 1121–1149.
Heckman, James J., Vytlacil, Edward J., 2001a. Local instrumental variables. In: Hsiao, Cheng, Morimune, Kimio, Powell, James L. (Eds.), Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya. In: International Symposia in Economic Theory and Econometrics, vol. 13. Cambridge University Press (Chapter 1).
Heckman, James J., Vytlacil, Edward J., 2001b. Policy-relevant treatment effects. The American Economic Review 91 (2), 107–111; Papers and Proceedings of the Hundred Thirteenth Annual Meeting of the American Economic Association.
Heckman, James J., Vytlacil, Edward J., 2005. Structural equations, treatment effects, and econometric policy evaluation. Econometrica 73 (3), 669–738.
Heckman, James J., Vytlacil, Edward J., 2007a. Econometric evaluation of social programs, part I: causal models, structural models and econometric policy evaluation. In: Heckman, J.J., Leamer, E.E., (Eds.), Handbook of Econometrics, first ed., vol. 6B, Elsevier (Chapter 70).
Heckman, James J., Vytlacil, Edward J., 2007b. Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments. In: Heckman, J.J., Leamer, E.E. (Eds.), Handbook of Econometrics. In: Handbook of Econometrics, vol. 6. Elsevier (Chapter 71).
Heckman, James J., Smith, Jeffrey A., 1998. Evaluating the Welfare State. NBER Working Papers 6542. National Bureau of Economic Research, Inc.
Heckman, James J., Robb Jr., Richard, 1985. Alternative methods for evaluating the impact of interventions. In: Heckman, James J., Singer, Burton (Eds.), Longitudinal Analysis of Labor Market Data. Cambridge University Press, New York.
Heckman, James J., Urzua, Sergio, 2009. Comparing IV with structural models: what simple IV can and cannot identify. Working Paper 14706, National Bureau of Economic Research. February.
Heckman, James J., Ichimura, H., Todd, Petra, 1998. Matching as an econometric evaluation estimator. Review of Economic Studies 65 (2), 261–294.
Heckman, James J., Tobias, Justin L., Vytlacil, Edward J., 2001. Four parameters of interest in the evaluation of social programs. Southern Economic Journal 68 (2), 210–223.
Heckman, James J., Tobias, Justin L., Vytlacil, Edward J., 2003. Simple estimators for treatment parameters in a latent variable framework. Review of Economics and Statistics 85 (3), 748–755.
Heckman, James J., LaLonde, Robert J., Smith, James A., 1999. The economics and econometrics of active labour market programmes. In: The Handbook of Labor Economics, vol. III. North–Holland, Amsterdam.
Heckman, James J., Urzua, Sergio, Vytlacil, Edward J., 2006. Understanding instrumental variables in models with essential heterogeneity. Review of Economics and Statistics 88 (3), 389–432.
Hirano, Keisuke, Imbens, Guido, Ridder, Geert, 2003. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71 (4), 1161–1189.
Holland, Paul W., 1986. Statistics and causal inference. Journal of the American Statistical Association 81 (396), 945–960.
Imbens, Guido, Angrist, Joshua, 1994. Identification and estimation of local average treatment effects. Econometrica 62 (2), 467–476.

Imbens, Guido, Lemieux, Thomas, 2008a. Regression discontinuity designs: a guide to practice. Journal of Econometrics 142 (2), 615–635.
Imbens, Guido, Lemieux, Thomas, 2008b. Regression discontinuity designs: a guide to practice. Journal of Econometrics.
Imbens, Guido W., 2009. Better LATE than nothing: some comments on Deaton (2009) and Heckman and Urzua (2009). Working Paper 14896, National Bureau of Economic Research, April.
Keane, Michael P., 2009. Structural vs. atheoretic approaches to econometrics. Journal of Econometrics, Corrected Proof.
Keane, Michael P., Wolpin, Kenneth I., 2007. Exploring the usefulness of a nonrandom holdout sample for model validation: welfare effects on female behavior. International Economic Review 48 (4), 1351–1378.
Lee, David S., 2008. Randomized experiments from non-random selection in US house elections. Journal of Econometrics 142 (2), 675–697.
Lee, David S., Card, David, 2008. Regression discontinuity inference with specification error. Journal of Econometrics 142 (2), 655–674. The regression discontinuity design: Theory and applications.
Lee, David S., Lemieux, Thomas, 2009. Regression discontinuity designs in economics. Working Paper 14723. National Bureau of Economic Research, February.
Lehmann, Erich Leo, 1959. Testing Statistical Hypotheses. John Wiley & Sons, Inc., New York.
Lehmann, Erich Leo, Hodges Jr., Joseph Lawson, 1964. Basic Concepts of Probability and Statistics. Holden-Day, San Francisco.
Lemieux, Thomas, Milligan, Kevin, 2008. Incentive effects of social assistance: a regression discontinuity approach. Journal of Econometrics 142 (2), 807–828.
Lucas Jr., Robert, 1976. Econometric policy evaluation: a critique. Carnegie-Rochester Conference Series on Public Policy 1 (1), 19–46.
Maddala, G.S., 1983. Limited-dependant and Qualitative Variables in Econometrics. Cambridge University Press.
Manning, Alan, 2003. Monopsony in Motion: Imperfect Competition in Labor Markets. Princeton University Press, Princeton, NJ.
Marschak, Jacob, 1953. Economic measurements for policy and prediction. In: Hood, William C., Koopmans, Tjalling C. (Eds.), Studies in Econometric Method, John Wiley and Sons, New York, pp. 1–26.
Mayo, Deborah G., 1996. Error and the growth of experimental knowledge. In: Science and Its Conceptual Foundations. University of Chicago Press, Chicago.
McCall, Brian Patrick, McCall, John Joseph, 2008. The economics of search, Routledge, London, New York.
McCrary, Justin, 2008. Manipulation of the running variable in the regression discontinuity design: a density test. Journal of Econometrics 142 (2), 698–714.
McCrary, Justin, Royer, Heather, 2010. The effect of female education on fertility and infant health: evidence from school entry laws using exact date of birth. Unpublished Working Paper. University of California Berkeley.
McFadden, Daniel, Talvitie, Antti, Associates. 1977. Demand Model Estimation and Validation, Urban Travel Demand Forecasting Project UCB-ITS-SR-77-9. The Institute of Transportation Studies, vol. V. University of California, Irvine and University of California, Berkeley. Phase 1 Final Report Series.
Millimet, Daniel L., Tchernis, Rusty, 2009. On the specification of propensity scores, with applications to the analysis of trade policies. Journal of Business and Economic Statistics 27 (3), 397–415.
Oreopoulos, Phillip, 2006. Estimating average and local average treatment effects of education when compulsory schooling laws really matter. American Economic Review 96 (1), 152–175.
Orr, Larry, Feins, Judith D., Jacob, Robin, Beecroft, Erik, Sanbonmatsu, Lisa, Katz, Lawrence F., Liebman, Jeffrey B., Kling, Jeffrey R., 2003. Moving to opportunity interim impacts evaluation. Final Report. US Department of Housing and Urban Development.
Pagan, A., Ullah, A., 1999. Nonparametric Econometrics. Cambridge University Press, New York.
Powell, James L., 1994. Estimation of semiparametric models. In: Engle, Robert, McFadden, Daniel (Eds.), Handbook of Econometrics, vol. 4. North Holland, Amsterdam.
Quandt, Richard E., 1958. The estimation of the parameters of a linear regression system obeying two separate regimes. Journal of the American Statistical Association 53 (284), 873–880.

Quandt, Richard E., 1972. A new approach to estimating switching regressions. Journal of the American Statistical Association 67 (338), 306–310.
Reiss, Peter C., Wolak, Frank A., 2007. Structural econometric modeling: rationales and examples from industrial organization. In: Heckman, James J., Leamer, E.E. (Eds.), Handbook of Econometrics. In: Handbook of Econometrics, vol. 6. Elsevier.
Robins, Philip K., 1985. A comparison of the labor supply findings from the four negative income tax experiments. The Journal of Human Resources 20 (4), 567–582.
Rosen, Sherwin, 1987. The theory of equalizing differences. In: Handbook of Labor Economics. In: Ashenfelter, O., Layard, R. (Eds.), Handbook of Labor Economics, vol. 1. Elsevier, pp. 641–692 (Chapter 12).
Rosenbaum, Paul, Rubin, Donald, 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70 (1), 41–55.
Rosenzweig, Mark R., Wolpin, Kenneth I., 2000. Natural 'natural experiments' in economics. Journal of Economic Literature 38 (4), 827–874.
Roy, A., 1951. Some thoughts on the distribution of earnings. Oxford Economic Papers 3 (2), 135–146.
Rubin, Donald B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66 (5), 688–701.
Rubin, Donald B., 1986. Statistics and causal inference: comment: which ifs have causal answers. Journal of the American Statistical Association 81 (396), 961–962.
Rubin, Donald B., 1990. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Comment: Neyman (1923) and causal inference in experiments and observational studies. Statistical Science 5 (4), 472–480.
Rubin, Donald B., Thomas, N., 1996. Matching using estimated propensity scores: relating theory to practice. Biometrics 52, 249.
Rust, John, 2009. Comments on: by Michael Keane. Journal of Econometrics, Corrected Proof.
Splawa-Neyman, Jerzy, Dabrowska, D.M., Speed, T.P., 1990. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Statistical Science 5 (4), 465–472.
Splawa-Neyman, Iwaszkiewicz, K., Kolodziejczyk, St., 1935. Statistical problems in agricultural experimentation. Supplement to the Journal of the Royal Statistical Society 2 (2), 107–180.
Taber, Christopher, French, Eric, 2011. Identification of models of the labor market. In: Ashenfelter, Orley, Card, David (Eds.), Handbook of Labor Economics, vol. 4A. Elsevier Science, pp. 537–617.
Tesfatsion, Leight, 2007. Introductory notes on complex adaptive systems and agent-based computational economics. Technical Report. Department of Economics. Iowa State University, January http://www.econ.iastate.edu/classes/econ308/tesfatsion/bat1a.htm.
Thistlethwaite, Donald L., Campbell, Donald T., 1960. Regression-discontinuity analysis: an alternative to the ex-post facto experiment. Journal of Educational Psychology 51, 309–317.
Todd, Petra, Wolpin, Kenneth, 2006. Ex ante evaluation of social programs. PIER Working Paper Archive. Penn Institute for Economic Research. Department of Economics. University of Pennsylvania.
Van der Klaauw, Wilbert, 2002. Estimating the effect of financial aid offers on college enrollment: a regression-discontinuity approach. International Economic Review 43 (4), 1249–1287.
Van der Klaauw, Wilbert, 2008a. Regression-discontinuity analysis: a survey of recent developments in economics. Labour 22 (2), 219–245.
Van der Klaauw, Wilbert, 2008b. Regression-discontinuity analysis: A survey of recent developments in economics. Labour 22 (2), 219–245.
Windrum, Paul, Fagiolo, Giorgio, Moneta, Alessio, 2007. Empirical validation of agent-based models: alternatives and prospects. Journal of Artificial Societies and Social Simulation 10.
Wolpin, Kenneth I., 2007. Ex ante policy evaluation, structural estimation and model selection. American Economic Review 97 (2), 48–52.

CHAPTER 6

Identification of Models of the Labor Market☆

Eric French[*], **Christopher Taber**[**]

[*] Federal Reserve Bank of Chicago

[**] Department of Economics, University of Wisconsin-Madison and NBER

Contents

☆ We thank Pedro Carneiro, Bruce Hansen, John Kennan, Salvador Navarro, Jim Walker, and students in Taber's 2010 Economics 751 class for comments. The opinions and conclusions are solely those of the authors, and should not be construed as representing the opinions of the Federal Reserve System. We thank Zach Seeskin and David Benson for excellent research assistance.

ISSN 0169-7218, DOI 10.1016/S0169-7218(11)00412-6

Abstract

This chapter discusses identification of common selection models of the labor market. We start with the classic Roy model and show how it can be identified with exclusion restrictions. We then extend the argument to the generalized Roy model, treatment effect models, duration models, search models, and dynamic discrete choice models. In all cases, key ingredients for identification are exclusion restrictions and support conditions.

JEL classification: C14; C51; J22; J24

Keywords: Identification; Roy model; Discrete choice; Selection; Treatment effects

1. INTRODUCTION

This chapter discusses identification of common selection models of the labor market. We are primarily concerned with nonparametric identification. We view nonparametric identification as important for the following reasons.

First, recent advances in computer power, more widespread use of large data sets, and better methods mean that estimation of increasingly flexible functional forms is possible. Flexible functional forms should be encouraged. The functional form and distributional assumptions used in much applied work rarely come from the theory. Instead, they come from convenience. Furthermore, they are often not innocuous.[1]

Second, the process of thinking about nonparametric identification is useful input into applied work. It is helpful to an applied researcher both in informing her about which type of data would be ideal and which aspects of the model she might have some hope of estimating. If a feature of the model is not nonparametrically identified, then one knows it cannot be identified directly from the data. Some additional type of functional form assumption must be made. As a result, readers of empirical papers are often skeptical of the results in cases in which the model is not nonparametrically identified.

Third, identification is an important part of a proof of consistency of a nonparametric estimator.

However, we acknowledge the following limitation of focusing on nonparametric identification. With any finite data set, an empirical researcher can almost never be completely nonparametric. Some aspects of the data that might be formally identified could never be estimated with any reasonable level of precision. Instead, estimators are usually only nonparametric in the sense that one allows the flexibility of the model to

[1] A classic reference on this is Lalonde (1986) who shows that parametric models cannot replicate the results of an experiment. Below we present an example on Catholic schools from Altonji et al. (2005a) suggesting that parametric assumptions drive the empirical estimates.

grow with the sample size. A nice example of this is Sieve estimators in which one estimates finite parameter models but the number of parameters gets large with the data set. An example would be approximating a function by a polynomial and letting the degree of the polynomial get large as the sample size increases. However, in that case one still must verify that the model is nonparametrically identified in order to show that the model is consistent. One must also construct standard errors appropriately. In this chapter we do not consider the purely statistical aspects of nonparametric estimation, such as calculation of standard errors. This is a very large topic within econometrics.[2]

The key issue in identification of most models of the labor market is the selection problem. For example, individuals are typically not randomly assigned to jobs. With this general goal in mind we begin with the simplest and most fundamental selection model in labor economics, the Roy (1951) model. We go into some detail to explain Heckman and Honoré's (1990) results on identification of this model. A nice aspect of identification of the Roy model is that the basic methodology used in this case can be extended to show identification of other labor models. We spend the rest of the chapter showing how this basic intuition can be used in a wide variety of labor market models. Specifically we cover identification in the generalized Roy model, treatment effect models, the competing risk model, search models, and forward looking dynamic models. While we are clearly not covering all models in labor economics, we hope the ideas are presented in a way that the similarities in the basic models can be seen and can be extended by the reader to alternative frameworks.

The plan of this chapter is specifically as follows. Section 2 discusses some econometric preliminaries. We consider the Roy model in Section 3, generalize this to the Generalized Roy model in Section 4, and then use the model to think about identification of treatment effects in Section 5. In Section 6 we consider duration models and search models and then consider estimation of dynamic discrete choice models in Section 7. Finally in Section 8 we offer some concluding thoughts.

2. ECONOMETRIC PRELIMINARIES

2.1. Notation

Throughout this chapter we use capital letters with i subscripts to denote random variables and small letters without i subscripts to denote possible outcomes of that random variable. We will also try to be explicit throughout this chapter in denoting conditioning. Thus, for example, we will use the notation

$$E(Y_i \mid X_i = x)$$

to denote the expected value of outcome Y_i conditional on the regressor variable X_i being equal to some realization x.

[2] See Chen (2007) for discussion of Sieve estimators, including standard error calculation.

2.2. Identification

The word "identification" has come to mean different things to different labor economists. Here, we use a formal econometrics definition of identification. Consider two different models that lead to two data generating processes. If the data generated by these two models have exactly the same distribution then the two models are not separately identified from each other. However, if any two different model specifications lead to different data distributions, the two specifications are separately identified. We give a more precise definition below. Our definition of identification is based on some of the notation and set up of Matzkin's (2007) following an exposition based on Shaikh (2010).

Let P denote the true distribution of the observed data X. An econometric model defines a data generating process. We assume that the model is specified up to an unknown vector θ of parameters, functions and distribution functions. This is known to lie in space Θ. Within the class of models, the element $\theta \in \Theta$ determines the distribution of the data that is observable to the researcher P_θ. Notice that identification is fundamentally data dependent. With a richer data set, the distribution P_θ would be a different object.

Let $\mathcal{P}$ be the set of all possible distributions that could be generated by the class of models we consider (i.e. $\mathcal{P} \equiv \{P_\theta : \theta \in \Theta\}$). We assume that the model is correctly specified, which means that $P \in \mathcal{P}$. The identified set is defined as

$$\Theta(P) \equiv \{\theta \in \Theta : P_\theta = P\}.$$

This is the set of possible θ that could have generated data that has distribution P. By assuming that $P \in \mathcal{P}$ we have assumed that our model is correctly specified so this set is not empty. We say that θ *is identified if* $\Theta(P)$ *is a singleton for all* $P \in \mathcal{P}$.

The question we seek to answer here is under what conditions is it possible to learn about θ (or some feature of θ) from the distribution of the observed data P. Our interest is not always to identify the full data generating process. Often we are interested in only a subset of the model, or a particular outcome from it. Specifically, our goal may be to identify

$$\psi = \Psi(\theta),$$

where Ψ is a known function. For example in a regression model $Y_i = X_i'\beta + u_i$, the feature of interest is typically the regression coefficients. In this case Ψ would take the trivial form

$$\Psi(\theta) = \beta.$$

However, this notation allows for more general cases in which we might be interested in identifying specific aspects of the model. For example, if our interest is in identifying the

covariance between X and Y in the case of the linear regression model, we do not need to know θ per se, but rather a transformation of these parameters. That is we could be interested in

$$\Psi(\theta) = \text{cov}(X_i, Y_i).$$

We could also be interested in a forecast of the model such as

$$\Psi(\theta) = x'\beta$$

for some specific x. The distinction between identification of features of the model as opposed to the full model is important, as in many cases the full model is not identified but the key feature of interest is identified.

To think about identification of ψ we define

$$\Psi(\Theta(P)) = \{\Psi(\theta) : \theta \in \Theta(P)\}.$$

That is, it is the set of possible values of ψ that are consistent with the data distribution P. We say that ψ *is identified if* $\Psi(\Theta(P))$ *is a singleton*.

As an example consider the standard regression model with two regressors:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \tag{2.1}$$

with $E(\varepsilon_i \mid X_i = x) = 0$ for any value x (where $X_i = (X_{1i}, X_{2i})$). In this case $\theta = (\beta, F_{X,\varepsilon})$, where $F_{X,\varepsilon}$ is the joint distribution of $(X_{1i}, X_{2i}, \varepsilon_i)$ and $\beta = (\beta_0, \beta_1, \beta_2)$. One would write Θ as $\mathcal{B} \times \mathcal{F}_{X,\varepsilon}$, where $\mathcal{B}$ is the parameter space for β and $\mathcal{F}_{X,\varepsilon}$ is the space of joint distributions between X_i and ε_i that satisfy $E(\varepsilon_i \mid X_i = x) = 0$ for all x. Since the data here is represented by (X_{1i}, X_{2i}, Y_i), P_θ represents the joint distribution of (X_{1i}, X_{2i}, Y_i). Given knowledge of β and $F_{X,\varepsilon}$ we know the data generating process and thus we know P_θ.

To focus ideas suppose we are interested in identifying β (i.e. $\Psi(\beta, F_{X,\varepsilon}) = \beta$) in regression model (2.1) above. Let the true value of the data generating process be $\theta^* = (\beta^*, F^*_{X,\varepsilon})$ so that by definition $P_{\theta^*} = P$. In this case $\Theta(P) \equiv \left\{(\beta, F_{X,\varepsilon}) \in \mathcal{B} \times \mathcal{F}_{X,\varepsilon} : P_{\beta, F_{x,\varepsilon}} = P\right\}$, that is it is the set of $(\beta, F_{X,\varepsilon})$ that would lead our data (X_i, Y_i) to have distribution P. In this case $\Psi(\Theta(P))$ is the set of values of β in this set (i.e. $\Psi(\Theta(P)) = \{\beta : (\beta, F_{X,\varepsilon}) \in \Theta(P)$ for some $F_{X,\varepsilon} \in \mathcal{F}_{X,\varepsilon}\}$).

In the case of 2 covariates, we know the model is identified as long as X_{1i} and X_{2i} are not degenerate and not collinear. To see how this definition of identification applies to this model, note that for any $\beta^* \neq \beta$ the lack of perfect multicollinearity means that

we can always find values of (x_1, x_2) for which

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 \neq \beta_0^* + \beta_1^* x_1 + \beta_2^* x_2.$$

Since $E(Y_i \mid X_i = x)$ is one aspect of the joint distribution of P_θ, it must be the case that when $\beta^* \neq \beta$, $P_\theta \neq P$. Since this is true for any value of $\beta \neq \beta^*$, then $\Psi(\Theta(P))$ must be the singleton β^*.

However, consider the well known case of perfect multicollinearity in which the model is not identified. In particular suppose that

$$X_{1i} + X_{2i} = 1.$$

For the true value of $\beta^* = (\beta_0^*, \beta_1^*, \beta_2^*)$ consider some other value $\widetilde{\beta} = (\beta_0^* + \beta_2^*, \beta_1^* - \beta_2^*, 0)$. Then for any x,

$$\begin{aligned} E(Y_i \mid X_i = x) &= \beta_0^* + \beta_1^* x_1 + \beta_2^* x_2 \\ &= \beta_0^* + \beta_1^* x_1 + \beta_2^* (1 - x_1) \\ &= \beta_0^* + \beta_2^* + \left(\beta_1^* - \beta_2^*\right) x_1 \\ &= \widetilde{\beta}_0 + \widetilde{\beta}_1 x_1. \end{aligned}$$

If $F_{X,\varepsilon}$ is the same for the two models, then the joint distribution of (Y_i, X_i) is the same in the two cases. Thus the identification condition above is violated because with $\widetilde{\theta} = (\widetilde{\beta}, F_{X,\varepsilon}^*)$, $P_{\widetilde{\theta}} = P$ and thus $\widetilde{\beta} \in \Psi(\Theta(P))$. Since the true value $\beta^* \in \Psi(\Theta(P))$ as well, $\Psi(\Theta(P))$ is not a singleton and thus β is not identified.

2.3. Support

Another important issue is the support of the data. The simplest definition of support is just the range of the data. When data are discrete, this is the set of values that occur with positive probability. Thus a binary variable that is either zero or one would have support $\{0, 1\}$. The result of a die roll has support $\{1, 2, 3, 4, 5, 6\}$. With continuous variables things get somewhat more complicated. One can think of the support of a random variable as the set of values for which the density is positive. For example, the support of a normal random variable would be the full real line (which we will often refer to as "full support"). The support of a uniform variable on $[0, 1]$ is $[0, 1]$. The support of an exponential variable would be the positive real line.

This can be somewhat trickier in dealing with outcomes that occur with measure zero. For example one could think of the support of a uniform variable as $[0, 1]$, $(0, 1]$, $[0, 1)$, or $(0, 1)$. The distinction between these objects will not be important in what we are doing, but to be formal we will use the Davidson (1994) definition of support. He defines the support of a random variable with distribution F

as the set of points at which F is (strictly) increasing.[3] By this definition, the support of a uniform would be $[0, 1]$. We will also use the notation $\text{supp}(Y_i)$ to denote the unconditional support of random variable Y_i and $\text{supp}(Y_i \mid X_i = x)$ to denote the conditional support.

To see the importance of this concept, consider a simple case of the separable regression model

$$Y_i = g(X_i) + u_i$$

with a single continuous X_i variable and $E(u_i \mid X_i = x) = 0$ for $x \in \text{supp}(X_i)$. In this case we know that

$$E(Y_i \mid X_i = x) = g(x).$$

Letting $\mathcal{X}$ be the support of X_i, it is straightforward to see that g is identified on the set $\mathcal{X}$. But g is not identified outside the set $\mathcal{X}$ because the data is completely silent about these values. Thus if $\mathcal{X} = \mathbb{R}$, g is globally identified. However, if $\mathcal{X}$ only covers a subset of the real line it is not. For example, one interesting counterfactual is the change in the expected value of Y_i if X_i were increased by $\delta : E(g(X_i + \delta))$. If $\mathcal{X} = \mathbb{R}$ this is trivially identified, but if the support of X_i were bounded from above, this would no longer be the case. That is, if the supremum of $\mathcal{X}$ is $\bar{x} < \infty$, then for any value of $x > \bar{x} - \delta$, $g(x + \delta)$ is not identified and thus the unconditional expected value of $g(X_i + \delta)$ is not identified either. This is just a restatement of the well known fact that one cannot project out of the data unless one makes functional form assumptions. Our point here is that support assumptions are very important in nonparametric identification results. One can only identify g over the range of plausible values of X_i if X_i has full support. For this reason, we will often make strong support condition assumptions. This also helps illuminate the tradeoff between functional form assumptions and flexibility. In order to project off the support of the data in a simple regression model one needs to use some functional form assumption. The same is true for selection models.

2.4. Continuity

There is one complication that we need to deal with throughout. It is not a terribly important issue, but will shape some of our assumptions. Consider again the separable regression model

$$Y_i = g(X_i) + u_i. \tag{2.2}$$

[3] He defines F (strictly) increasing at point x to mean that for any $\varepsilon > 0, F(x + \varepsilon) > F(x - \varepsilon)$.

As mentioned above $E(Y_i \mid X_i = x) = g(x)$, so it seems trivial to see that g is identified, but that is not quite true. To see the problem, suppose that both X_i and u_i are standard normals. Consider two different models for g,

Model 1:

$$g(x) = \begin{cases} 0 & x < 1.4 \\ 1 & x \geq 1.4 \end{cases}$$

versus

Model 2:

$$g(x) = \begin{cases} 0 & x \leq 1.4 \\ 1 & x > 1.4. \end{cases}$$

These models only differ at the point $x = 1.4$, but since X_i is normal this is a zero probability event and we could never distinguish between these models because they imply the same joint distribution of (X_i, Y_i). For the exact same reason it isn't really a concern (except in very special cases such as if one was evaluating a policy in which we would set $X_i = 1.4$ for everyone). Since this will be an issue throughout this chapter we explain how to deal with it now and use this convention throughout the chapter.

We will make the following assumptions.

Assumption 2.1. X_i can be written as (X_i^c, X_i^d), where the elements of X_i^c are continuously distributed (no point has positive mass), and X_i^d is distributed discretely (all support points have positive mass).

Assumption 2.2. For any $x^d \in \text{supp}(X_i^d)$, $g(x^c, x^d)$ is almost surely continuous across $x^c \in \text{supp}(X_i^c \mid X_i^d = x^d)$.

The first part says that we can partition our observables into continuous and discrete ones. One could easily allow for variables that are partially continuous and partially discrete, but this would just make our results more tedious to exposit. The second assumption states that choosing a value of X at which g is discontinuous (in the continuous variables) is a zero probability event.

Theorem 2.1. *Under Assumptions 2.1 and 2.2 and assuming model (2.2) with* $E(u_i \mid X_i = x) = 0$ *for* $x \in \text{supp}(X_i)$, $g(x)$ *is identified on a set* $\mathcal{X}^*$ *that has measure* 1.

(Proof in Appendix.)

The proof just states that g is identified almost everywhere. More specifically it is identified everywhere that it is continuous.

3. THE ROY MODEL

The classic model of selection in the labor market is the Roy (1951) model. In the Roy model, workers choose one of two possible occupations: hunting and fishing. They cannot pursue both at the same time. The worker's log wage is Y_{fi} if he fishes and Y_{hi} if he hunts. Workers maximize income so they choose the occupation with the higher wage. Thus a worker chooses to fish if $Y_{fi} > Y_{hi}$. The occupation is defined as

$$J_i = \begin{cases} f & \text{if } Y_{fi} > Y_{hi} \\ h & \text{if } Y_{hi} \geq Y_{fi} \end{cases} \tag{3.1}$$

and the log wage is defined as

$$Y_i = \max\{Y_{fi}, Y_{hi}\}. \tag{3.2}$$

Workers face a simple binary choice: choose the job with the highest wage. This simplicity has led the model to be used in one form or another in a number of important labor market contexts. Many discrete choice models share the Roy model's structure. Examples in labor economics include the choice of whether to continue schooling, what school to attend, what occupation to pursue, whether to join a union, whether to migrate, whether to work, whether to obtain training, and whether to marry.

As mentioned in the introduction, we devote considerable attention to identification of this model. In subsequent sections we generalize these results to other models.

The responsiveness of the supply of fishermen to changes in the price of fish depends critically on the joint distribution of (Y_{fi}, Y_{hi}). Thus we need to know what a fisherman would have made if he had chosen to hunt. However, we do not observe this but must infer its counterfactual distribution from the data at hand. Our focus is on this selection problem. Specifically, much of this chapter is concerned with the following question: **Under what conditions is the joint distribution of (Y_{fi}, Y_{hi}) identified?** We start by considering estimation in a parametric model and then consider nonparametric identification.

Roy (1951) is concerned with how occupational choice affects the aggregate distribution of earnings and makes a series of claims about this relationship. These claims turn out to be true when the distribution of skills in the two occupations is lognormal.

Heckman and Honoré (1990) consider identification of the Roy model (i.e., the joint distribution of (Y_{fi}, Y_{hi})). They show that there are two methods for identifying the Roy model. The first is through distributional assumptions. The second is through exclusion restrictions.[4]

[4] Heckman and Honoré discuss price variation as separate from exclusion restrictions. However, in our framework price changes can be modeled as just one type of exclusion restriction so we do not explicitly discuss price variation.

In order to focus ideas, we use the following case:

$$Y_{fi} = g_f(X_{fi}, X_{0i}) + \varepsilon_{fi} \tag{3.3}$$
$$Y_{hi} = g_h(X_{hi}, X_{0i}) + \varepsilon_{hi}, \tag{3.4}$$

where the unobservable error terms $(\varepsilon_{fi}, \varepsilon_{hi})$ are independent of the observable variables $X_i = (X_{fi}, X_{hi}, X_{0i})$ and Y_{fi} and Y_{hi} denote log wages in the fishing and hunting sectors respectively. We distinguish between three types of variables. X_{0i} influences productivity in both fishing and hunting, X_{fi} influences fishing only, and X_{hi} influences hunting only. The variables X_{fi} and X_{hi} are "exclusion restrictions," and play a very important role in the identification results below. In the context of the Roy model, an exclusion restriction could be a change in the price of rabbits which increases income from hunting, but not from fishing. The notation is general enough to incorporate a model without exclusion restrictions (in which case one or more of the X_{ji} would be empty).

Our version of the Roy framework imposes two strong assumptions. First, that Y_{ji} is separable in $g_j(X_{ji}, X_{0i})$ and ε_{ji} for $j \in \{f, h\}$. Second, we assume that $g_j(X_{ji}, X_{0i})$ and ε_{ji} are independent of one another. Note that independence implies homoskedasticity: the variance of ε_{ji} cannot depend on X_{ji}. There is a large literature looking at various other more flexible specifications and this is discussed thoroughly in Matzkin (2007). It is also trivial to extend this model to allow for a general relationship between X_{0i} and $(\varepsilon_{fi}, \varepsilon_{hi})$, as we discuss in Section 3.3 below.

We focus on the separable independent model for two reasons. First, the assumptions of separability and independence have bite beyond a completely general nonparametric relationship. That is, to the extent that they are true, identification is facilitated by these assumptions. Presumably because researchers think these assumptions are approximately true, virtually all empirical research uses these assumptions. Second, despite these strong assumptions, they are obviously much weaker than the standard assumptions that g is linear (i.e. $g_f(X_{fi}, X_{0i}) = X'_{fi}\gamma_{ff} + X'_{0i}\gamma_{0f}$) and that ε_{fi} is normally distributed. One approach to writing this chapter would have been to go through all of the many specifications and alternative assumptions. We choose to focus on a single base specification for expositional simplicity.

Heckman and Honoré (1990) first discuss identification of the joint distribution of $\left(Y_{fi}, Y_{hi}\right)$ using distributional assumptions. They show that when one can observe the distribution of wages in both sectors, and assuming $\left(Y_{fi}, Y_{hi}\right)$ is joint normally distributed, then the joint distribution of $\left(Y_{fi}, Y_{hi}\right)$ is identified from a single cross section even without any exclusion restrictions or regressors. To see why, write equations (3.3) and (3.4) without regressors (so $g_f = \mu_f$, the mean of Y_{fi}):

$$Y_{fi} = \mu_f + \varepsilon_{fi}$$
$$Y_{hi} = \mu_h + \varepsilon_{hi}$$

where

$$\begin{bmatrix} \varepsilon_{fi} \\ \varepsilon_{hi} \end{bmatrix} = N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_f^2 & \sigma_{fh} \\ \sigma_{fh} & \sigma_h^2 \end{bmatrix} \right).$$

Letting

$$\lambda(\cdot) = \frac{\phi(\cdot)}{\Phi(\cdot)}$$

(with ϕ and Φ the pdf and cdf of a standard normal),

$$c = \frac{\mu_f - \mu_h}{\sqrt{\sigma_f^2 + \sigma_h^2 - 2\sigma_{fh}}},$$

and for each $j \in \{h, f\}$,

$$\tau_j = \frac{\sigma_j^2 - \sigma_{fh}}{\sqrt{\sigma_f^2 + \sigma_h^2 - 2\sigma_{fh}}}.$$

One can derive the following conditions from properties of normal random variables found in Heckman and Honoré (1990):

$$\begin{aligned}
\Pr(J_i = f) &= \Phi(c) \\
E(Y_i \mid J_i = f) &= \mu_f + \tau_f \lambda(c) \\
E(Y_i \mid J_i = h) &= \mu_h + \tau_h \lambda(-c) \\
\operatorname{var}(Y_i \mid J_i = f) &= \sigma_f^2 + \tau_f^2(-\lambda(c)\, c - \lambda^2(c)) \\
\operatorname{var}(Y_i \mid J_i = h) &= \sigma_h^2 + \tau_h^2(\lambda(-c)\, c - \lambda^2(-c)) \\
E([Y_i - E(Y_i \mid J_i = f)]^3 \mid J_i = f) &= \tau_f^3 \lambda(c)[2\lambda^2(c) + 3c\lambda(c) + c^2 - 1] \\
E([Y_i - E(Y_i \mid J_i = h)]^3 \mid J_i = h) &= \tau_h^3 \lambda(-c)[2\lambda^2(-c) - 3c\lambda(-c) + c^2 - 1].
\end{aligned}$$

This gives us seven equations in the five unknowns $\mu_f, \mu_h, \sigma_f^2, \sigma_h^2$, and σ_{fh}. It is straightforward to show that the five parameters can be identified from this system of equations.

However, Theorems 7 and 8 of Heckman and Honoré (1990) show that when one relaxes the log normality assumption, without exclusion restrictions in the outcome

equation, the model is no longer identified. This is true despite the strong assumption of agent income maximization. This result is not particularly surprising in the sense that our goal is to estimate a full joint distribution of a two dimensional object (Y_{fi}, Y_{hi}), but all we can observe is two one dimensional distributions (wages conditional on job choice). Since there is no information in the data about the wage that a fisherman may have received as a hunter, one cannot identify this joint distribution. In fact, Theorem 7 of Heckman and Honoré (1990) states that we can never distinguish the actual model from an alternative model in which skills are independent of each other.

3.1. Estimation of the normal linear labor supply model

It is often the case that we only observe wages in one sector. For example, when estimating models of participation in the labor force, the wage is observed only if the individual works. We can map this into our model by associating working with "fishing" and not working with "hunting." That is, we let Y_{fi} denote income if working and let Y_{hi} denote the value of not working.[5]

But there are other examples in which we observe the wage in only one sector. For example, in many data sets we do not observe wages of workers in the black market sector. Another example is return immigration in which we know when a worker leaves the data to return to their home country, but we do not observe that wage.

In Section 3.2 we discuss identification of the nonparametric version of the model. However, it turns out that identification of the more complicated model is quite similar to estimation of the model with normally distributed errors. Thus we review this in detail before discussing the nonparametric model. We also remark that providing a consistent estimator also provides a constructive proof of identification, so one can also interpret these results as (informally) showing identification in the normal model. The model is similar to Willis and Rosen's (1979) Roy Model of educational choices or Lee's (1978) model of union status and the empirical approach is analogous. We assume that

$$\begin{aligned} Y_{fi} &= X_{fi}'\gamma_{ff} + X_{0i}'\gamma_{0f} + \varepsilon_{fi} \\ Y_{hi} &= X_{hi}'\gamma_{hh} + X_{0i}'\gamma_{0h} + \varepsilon_{hi} \\ \begin{bmatrix} \varepsilon_{fi} \\ \varepsilon_{hi} \end{bmatrix} &= N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_f^2 & \sigma_{fh} \\ \sigma_{fh} & \sigma_h^2 \end{bmatrix} \right). \end{aligned}$$

In a labor supply model where f represents market work, Y_{fi} is the market wage which will be observed for workers only. Y_{hi}, the pecuniary value of not working, is never

[5] There are two common participation models. The first is the home production model in which the individual chooses between home and market production. The second is the labor supply model in which the individual chooses between market production and leisure. In practice the two types of models tend to be similar and some might argue the distinction is semantic. In a model of home production, Y_{hi} is the (unobserved) gain from home production. In a model of labor supply, Y_{hi} is the leisure value of not working.

observed in the data. Keane et al. (2011) example of the static model of a married woman's labor force participation is similar.

One could simply estimate this model by maximum likelihood. However we discuss a more traditional four step method to illustrate how the parametric model is identified. This four step process will be analogous to the more complicated nonparametric identification below. Step 1 is a "reduced form probit" of occupational choices as a function of all covariates in the model. Step 2 estimates the wage equations by controlling for selection as in the second step of a Heckman Two step (Heckman, 1979). Step 3 uses the coefficients of the wage equations and plugs these back into a probit equation to estimate a "structural probit." Step 4 shows identification of the remaining elements of the variance-covariance matrix of the residuals.

Step 1: Estimation of choice model

The probability of choosing fishing (i.e., work) is:

$$\begin{aligned}
\Pr\left(J_i = f \mid X_i = x\right) &= \Pr\left(Y_{fi} > Y_{hi} \mid X_i = x\right) \\
&= \Pr\left(x_f'\gamma_{ff} + x_0'\gamma_{0f} + \varepsilon_{fi} > x_0'\gamma_{0h} + x_h'\gamma_{hh} + \varepsilon_{hi}\right) \\
&= \Pr\left(x_f'\gamma_{ff} - x_h'\gamma_{hh} + x_0'\left(\gamma_{0f} - \gamma_{0h}\right) > \varepsilon_{hi} - \varepsilon_{fi}\right) \\
&= \Phi\left(\frac{x_f'\gamma_{ff} - x_h'\gamma_{hh} + x_0'\left(\gamma_{0f} - \gamma_{0h}\right)}{\sigma^*}\right) \\
&= \Phi\left(x'\gamma^*\right) \qquad (3.5)
\end{aligned}$$

where Φ is the cdf of a standard normal, σ^* is the standard deviation of $\left(\varepsilon_{hi} - \varepsilon_{fi}\right)$, and

$$\gamma^* \equiv \left(\frac{\gamma_{ff}}{\sigma^*}, \frac{-\gamma_{hh}}{\sigma^*}, \frac{\gamma_{0f} - \gamma_{0h}}{\sigma^*}\right).$$

This is referred to as the "reduced form model" as it is a reduced form in the classical sense: the parameters are a known function of the underlying structural parameters. It can be estimated by maximum likelihood as a probit model. Let $\widehat{\gamma^*}$ represent the estimated parameter vector. This is all that can be learned from the choice data alone. We need further information to identify σ^* and to separate γ_{0f} from γ_{0h}.

Step 2: Estimating the wage equation

This is essentially the second stage of a Heckman (1979) two step. To review the idea behind it, let

$$\varepsilon_i^* = \frac{\varepsilon_{hi} - \varepsilon_{fi}}{\sigma^*}.$$

Then consider the regression

$$\varepsilon_{fi} = \tau \varepsilon_i^* + \zeta_i$$

where $\text{cov}\left(\varepsilon_i^*, \zeta_i\right) = 0$ (by definition of regression) and thus:

$$\begin{aligned}\tau &= \frac{\text{cov}\left(\varepsilon_{fi}, \varepsilon_i^*\right)}{\text{var}\left(\varepsilon_i^*\right)} \\ &= E\left[\varepsilon_{fi}\left(\frac{\varepsilon_{hi} - \varepsilon_{fi}}{\sigma^*}\right)\right] \\ &= \frac{\sigma_{fh} - \sigma_f^2}{\sigma^*}.\end{aligned}$$

The wage of those who choose to work is

$$\begin{aligned}E\left(Y_{fi} \mid J_i = f, X_i = x\right) &= x_f'\gamma_{ff} + x_0'\gamma_{0f} + E\left(\varepsilon_{fi} \mid J_i = f, X_i = x\right) \\ &= x_f'\gamma_{ff} + x_0'\gamma_{0f} + E\left(\tau\varepsilon_i^* + \zeta_i \mid \varepsilon_i^* \le x'\gamma^*\right) \\ &= x_f'\gamma_{ff} + x_0'\gamma_{0f} + \tau E\left(\varepsilon_i^* \mid \varepsilon_i^* \le x'\gamma^*\right) \\ &= x_f'\gamma_{ff} + x_0'\gamma_{0f} - \tau\lambda\left(x'\gamma^*\right). \qquad (3.6)\end{aligned}$$

Showing that $E\left(\varepsilon_i^* \mid \varepsilon_i^* \le x'\gamma^*\right) = -\lambda\left(x'\gamma^*\right)$ is a fairly straightforward integration problem and is well known. Because Eq. (3.6) is a conditional expectation function, OLS regression of Y_i on X_{0i}, X_{fi}, and $\lambda\left(X_i'\widehat{\gamma^*}\right)$ gives consistent estimates of γ_{ff}, γ_{0f}, and τ. $\widehat{\gamma^*}$ is the value of γ^* estimated in Eq. (3.5).

Note that we do not require an exclusion restriction. Since λ is a nonlinear function, but g_f is linear, this model is identified. However, without an exclusion restriction, identification is purely through functional form. When we consider a nonparametric version of the model below, exclusion restrictions are necessary. We discuss this issue in Section 3.2.

Step 3: The structural probit

Our next goal is to estimate γ_{0h} and γ_{hh}. In Step 1 we obtained consistent estimates of $\gamma^* \equiv \left(\frac{\gamma_{0f} - \gamma_{0h}}{\sigma^*}, \frac{\gamma_{ff}}{\sigma^*}, \frac{-\gamma_{hh}}{\sigma^*}\right)$ and in Step 2 we obtained consistent estimates of γ_{0f} and γ_{ff}.

When there is only one exclusion restriction (i.e. γ_{ff} is a scalar), identification proceeds as follows. Because we identified γ_{ff} in Step 2 and γ_{ff}/σ^* in Step 1, we can identify σ^*. Once σ^* is identified, it is easy to see how to identify γ_{hh} (because $\frac{-\gamma_{hh}}{\sigma^*}$ is identified) and γ_{0h} (because $\frac{\gamma_{0f} - \gamma_{0h}}{\sigma^*}$ and γ_{0f} are identified).

In terms of estimation of these objects, if there is more than one exclusion restriction the model is over-identified. If we have two exclusion restrictions, γ_{ff} and γ_{ff}/σ^* are both 2×1 vectors, and thus we wind up with 2 consistent estimates of σ^*. The most standard way of solving this model is by estimating the "structural probit:"

$$\Pr(J_i = f \mid X_i = x) = \Phi\left(\frac{1}{\sigma^*}\left(x_f'\widehat{\gamma_{ff}} + x_0'\widehat{\gamma_{0f}}\right) - x_h'\frac{\gamma_{hh}}{\sigma^*} - x_0'\frac{\gamma_{0h}}{\sigma^*}\right). \quad (3.7)$$

That is, one just runs a probit of J_i on $(X_{fi}'\widehat{\gamma_{ff}} + X_{0i}'\widehat{\gamma_{0f}})$, X_{0i}, and X_{hi} where $\widehat{\gamma_{ff}}$ and $\widehat{\gamma_{0f}}$ are our estimates of γ_{ff} and γ_{0f}.

Step 3 is essential if our goal is to estimate the labor supply equation. If we are only interested in controlling for selection to obtain consistent estimates of the wage equation, we do not need to worry about the structural probit. However, notice that

$$\frac{\partial \Pr(J_i = f \mid X_i = x)}{\partial Y_{fi}} = \frac{1}{\sigma^*}\phi\left(x'\gamma^*\right).$$

and thus the labor supply elasticity is:

$$\begin{aligned}\frac{\partial \log[\Pr(J_i = f \mid X_i = x)]}{\partial Y_{fi}} &= \frac{\partial \Pr(J_i = f \mid X_i = x)}{\partial Y_{fi}}\frac{1}{\Pr(J_i = f \mid X_i = x)} \\ &= \frac{1}{\sigma^*}\frac{\phi\left(x'\gamma^*\right)}{\Phi\left(x'\gamma^*\right)},\end{aligned}$$

where, as before, Y_{fi} is the log of income if working. Thus knowledge of σ^* is essential for identifying the effects of wages on participation.

One could not estimate the structural probit without the exclusion restriction X_{fi} as the first two components of the probit in Eq. (3.7) would be perfectly collinear. For any $\sigma^* > 0$ we could find a value of γ_{0h} and γ_{hh} that delivers the same choice probabilities. Furthermore, if these parameters were not identified, the elasticity of labor supply with respect to wages would not be identified either.

Step 4: Estimation of the variance matrix of the residuals

Lastly, we identify all the components of Σ, $(\sigma_f^2, \sigma_h^2, \sigma_{fh})$ as follows. We have described how to obtain consistent estimates of $\sigma^* = \sqrt{\sigma_f^2 + \sigma_h^2 - 2\sigma_{fh}}$ and $\tau = \frac{\sigma_{fh} - \sigma_f^2}{\sigma^*}$. This gives us two equations in three parameters. We can obtain the final equation by using the variance of the residual in the selection model since

$$\operatorname{var}(\varepsilon_{fi} \mid J_i = f, X_i = x) = \sigma_f^2 + \tau^2\left[-\lambda(x'\gamma^*)x'\gamma^* - \lambda^2(x'\gamma^*)\right].$$

Let $i = 1, \ldots, N_f$ index the set of individuals who choose $J_i = f$ and $\widehat{\varepsilon_{fi}}$ is the residual $Y_{fi} - X'_{fi}\widehat{\gamma_{ff}} - X'_{0i}\widehat{\gamma_{0f}}$ for individuals who choose $J_i = f$. Using "hats" to denote estimators we can estimate the model as

$$\widehat{\sigma_f^2} = \frac{1}{N_f}\sum_{i=1}^{N_f}\left(\widehat{\varepsilon_{fi}} + \tau\lambda\left(X_i'\widehat{\gamma^*}\right)\right)^2 - \widehat{\tau}^2\left(-\lambda\left(X_i'\widehat{\gamma^*}\right)X_i'\widehat{\gamma^*} - \lambda^2\left(X_i'\widehat{\gamma^*}\right)\right)$$
$$\widehat{\sigma_{fh}} = \widehat{\sigma_f^2} - \widehat{\tau}\widehat{\sigma^*}$$
$$\widehat{\sigma_h}^2 = \widehat{\sigma^*}^2 - \widehat{\sigma_f}^2 + 2\widehat{\sigma_{fh}}.$$

3.2. Identification of the Roy model: the non-parametric approach

Although the parametric case with exclusion restrictions is more commonly known, the model in the previous section is still identified non-parametrically if the researcher is willing to impose stronger support conditions on the observable variables. Heckman and Honoré (1990, Theorem 12) provide conditions under which one can identify the model nonparametrically using exclusion restrictions. We present this case below.

Assumption 3.1. $(\varepsilon_{fi}, \varepsilon_{hi})$ is continuously distributed with distribution function G, support $\mathbb{R}^2$, and is independent of X_i. The marginal distributions of ε_{fi} and $\varepsilon_{fi} - \varepsilon_{hi}$ have medians equal to zero.

Assumption 3.2. $\text{supp}(g_f(X_{fi}, x_0), g_h(X_{hi}, x_0)) = \mathbb{R}^2$ for all $x_0 \in \text{supp}(X_{0i})$.

Assumption 3.2 is crucial for identification. It states that for any value of $g_h(x_h, x_0)$, $g_f(X_{fi}, x_0)$ varies across the full real line and for any value of $g_f(x_f, x_0)$, $g_h(X_{hi}, x_0)$ varies across the full real line. This means that we can condition on a set of variables for which the probability of being a hunter (i.e. $\Pr(J_i = h|X_i = x)$) is arbitrarily close to 1. This is clearly a very strong assumption that we will discuss further.

We need the following two assumptions for the reasons discussed in Section 2.4.

Assumption 3.3. $X_i = (X_{fi}, X_{hi}, X_{0i})$ can be written as $(X_{fi}^c, X_{fi}^d, X_{hi}^c, X_{hi}^d, X_{0i}^c, X_{0i}^d)$ where the elements of $(X_{fi}^c, X_{hi}^c, X_{0i}^c)$ are continuously distributed (no point has positive mass), and $(X_{fi}^d, X_{hi}^d, X_{0i}^d)$ is distributed discretely (all support points have positive mass).

Assumption 3.4. For any $(x_f^d, x_h^d, x_0^d) \in \text{supp}(X_{fi}^d, X_{hi}^d, X_{0i}^d)$, $g_f(x_f^c, x_f^d, x_0^c, x_0^d)$ and $g_h(x_h^c, x_h^d, x_0^c, x_0^d)$ are almost surely continuous across $x^c \in \text{supp}(X_i^c \mid X_i^d = x^d)$.

Under these assumptions we can prove the theorem following Heckman and Honoré (1990).

Theorem 3.1. *If $(J_i \in \{f, h\}, Y_{fi}$ if $J_i = f, X_i)$ are all observed and generated under model (3.1)–(3.4), then under Assumptions 3.1–3.4, g_f, g_h, and G are identified on a set $\mathcal{X}^*$ that has measure 1.*

(Proof in Appendix.)

A key theme of this chapter is that the basic structure of identification in this model is similar to identification of more general selection models, so we explain this result in much detail. The basic structure of the proof we present below is similar to Heckman and Honoré's proof of their Theorems 10 and 12. We modify the proof to allow for the case where Y_{hi} is not observed.

The proof in the Appendix is more precise, but in the text we present the basic ideas. We follow a structure analogous to the parametric empirical approach when the residuals are normally distributed as presented in Section 3.1. First we consider identification of the occupational choice given only observable covariates and the choice model. This is the nonparametric analogue of the reduced form probit. Second we estimate g_f given the data on Y_{fi}, which is the analogue of the second stage of the Heckman two step, and is more broadly the nonparametric version of the classical selection model. In the third step we consider the nonparametric analogue of identification of the structural probit. Since we will have already established identification of g_f, identification of this part of the model boils down to identification of g_h. Finally in the fourth step we consider identification of G (the joint distribution of $(\varepsilon_{fi}, \varepsilon_{hi})$). We discuss each of these steps in order.

To map the Roy model into our formal definition of identification presented in Section 2.2, the model is determined by $\theta = (g_f, g_h, G, F_x)$, where F_x is the joint distribution of (X_{fi}, X_{hi}, X_{0i}). The observable data here is $(X_{fi}, X_{hi}, X_{0i}, J_i, 1(J_i = f)Y_{fi})$. Thus P is the joint distribution of this observable data and $\Theta(P)$ represents the possible data generating processes consistent with P.

Step 1: Identification of choice model

The nonparametric identification of this model is established in Matzkin (1992). We can write the model as

$$\begin{aligned}\Pr(J_i = f \mid X_i = x) &= \Pr(\varepsilon_{hi} - \varepsilon_{fi} < g_f(x_f, x_0) - g_h(x_h, x_0)) \\ &= G_{h-f}(g_f(x_f, x_0) - g_h(x_h, x_0)),\end{aligned}$$

where G_{h-f} is the distribution function of $\varepsilon_{hi} - \varepsilon_{fi}$.

Using data only on choices, this model is only identified up to a monotonic transformation. To see why, note that we can write $J_i = f$ when

$$g_f(x_f, x_0) - g_h(x_h, x_0) > \varepsilon_{hi} - \varepsilon_{fi} \tag{3.8}$$

but this is equivalent to the condition

$$M(g_f(x_f, x_0) - g_h(x_h, x_0)) > M(\varepsilon_{hi} - \varepsilon_{fi}) \tag{3.9}$$

where $M(.)$ is any strictly increasing function. Clearly the model in Eq. (3.8) cannot

be distinguished from an alternative model in Eq. (3.9). This is the nonparametric analog of the problem that the scale (i.e., the variance of $\varepsilon_{hi} - \varepsilon_{fi}$) and location (only the difference between $g_f(x_f, x_0)$ and $g_h(x_h, x_0)$ but not the level of either) of the parametric binary choice model are not identified. Without loss of generality we can normalize the model up to a monotonic transformation. There are many ways to do this. A very convenient normalization is to choose the transformation $M(\cdot) = G_{h-f}(\cdot)$ because $G_{h-f}\left(\varepsilon_{hi} - \varepsilon_{fi}\right)$ has a uniform distribution.[6] So we define

$$\begin{aligned} \varepsilon_i &\equiv G_{h-f}(\varepsilon_{hi} - \varepsilon_{fi}) \\ g(x) &\equiv G_{h-f}(g_f(x_f, x_0) - g_h(x_h, x_0)). \end{aligned}$$

Then

$$\begin{aligned} \Pr(J_i = f \mid X_i = x) &= \Pr(g_f(x_f, x_0) - g_h(x_h, x_0) > \varepsilon_{hi} - \varepsilon_{fi}) \\ &= \Pr(G_{h-f}(g_f(x_f, x_0) - g_h(x_h, x_0)) > G_{h-f}(\varepsilon_{hi} - \varepsilon_{fi})) \\ &= \Pr(\varepsilon_i < g(x)) \\ &= g(x). \end{aligned}$$

Thus we have established that we can (i) write the model as $J_i = f$ if and only if $g(X_i) > \varepsilon_i$ where ε_i is uniform $[0, 1]$ and (ii) that g is identified.

This argument can be mapped into our formal definition of identification from Section 2.2 above. The goal here is identification of g, so we define $\Psi(\theta) = g$. Note that even though g is not part of θ, it is a known function of the components of θ. The key set now is $\Psi(\Theta(P))$, which is now defined as the set of possible values g that could have generated the joint distribution of $(X_{fi}, X_{hi}, X_{0i}, J_i, 1(J_i = f)Y_{fi})$. Since $\Pr(J_i = f \mid X_i = x) = g(x)$, no other possible value of g could generate the data. Thus $\Psi(\Theta(P))$ only contains the true value and is thus a singleton.

Step 2: Identification of the wage equation g_f

Next consider identification of g_f. Median regression identifies

$$\text{Med}(Y_i \mid X_i = x, J_i = f) = g_f(x_f, x_0) + \text{Med}(\varepsilon_{fi} \mid X_i = x, \varepsilon_i < g(x)).$$

The goal is to identify $g_f(x_f, x_0)$. The problem is that when we vary (x_f, x_0) we also typically vary $\text{Med}(\varepsilon_{fi} \mid X_i = x, g(x) > \varepsilon_i)$. This is the standard selection problem. Because we can add any constant to g_f and subtract it from ε_{fi} without changing the model, a normalization that allows us to pin down the location of g_f is that $\text{Med}(\varepsilon_{fi}) = 0$. The problem is that this is the unconditional median rather than

[6] To see why note that for any x, $\Pr(G_{h-f}\left(\varepsilon_{hi} - \varepsilon_{fi}\right) < x) = \Pr(\varepsilon_{hi} - \varepsilon_{fi} \leq G^{-1}_{h-f}(x)) = G_{h-f}(G^{-1}_{h-f}(x)) = x$.

the conditional one. The solution here is what is often referred to as identification at infinity (e.g. Chamberlain, 1986, or Heckman, 1990). For some value (x_f, x_0) suppose we can find a value of x_h to send $\Pr(\varepsilon_i < g(x))$ arbitrarily close to one. It is referred to as identification at infinity because if g_h were linear in the exclusion restriction x_h this could be achieved by sending $x_h \to -\infty$. In our fishing/hunting example, this could be sending the price of rabbits to zero which in turn sends log income from hunting to $-\infty$. Then notice that[7]

$$\begin{aligned}\lim_{g(x)\to 1} \text{Med}(Y_i \mid X_i = x, J_i = f) &= g_f(x_f, x_0) + \lim_{g(x)\to 1} \text{Med}(\varepsilon_{fi} \mid \varepsilon_i \leq g(x)) \\ &= g_f(x_f, x_0) + \text{Med}(\varepsilon_{fi} \mid \varepsilon_i \leq 1) \\ &= g_f(x_f, x_0) + \text{Med}(\varepsilon_{fi}) \\ &= g_f(x_f, x_0).\end{aligned}$$

Thus g_f is identified.

Conditioning on x so that $\Pr(J_i = 1 \mid X_i = x)$ is arbitrarily close to one is essentially conditioning on a group of individuals for whom there is no selection, and thus there is no selection problem. Thus we are essentially saying that if we can condition on a group of people for whom there is no selection we can solve the selection bias problem.

While this may seem like cheating, without strong functional form assumptions it is necessary for identification. To see why, suppose there is some upper bound of $\text{supp}[g(X_i)]$ equal to $g^u < 1$ which would prevent us from using this type of argument. Consider any potential worker with a value of $\varepsilon_i > g^u$. For those individuals it must be the case that

$$\varepsilon_i > g(X_i)$$

so they must always be a hunter. As a result, the data is completely uninformative about the distribution of ε_{fi} for these individuals. For this reason the unconditional median of ε_{fi} would not be identified. We will discuss approaches to dealing with this potential problem in the Treatment Effect section below.

To relate this to the framework from Section 2.2 above now we define $\Psi(\theta) = g_f$, so $\Psi(\Theta(P))$ contains the values of g_f consistent with P. However since

$$\lim_{g(x)\to\infty} \text{Med}(Y_f \mid X_i = x, J_i = f) = g_f(x_f, x_0),$$

g_f is the only element of $\Psi(\Theta(P))$, thus it is identified.

[7] We are using loose notation here. What we mean by $lim_{g(x)\to 1}$ is to hold (x_f, x_0) fixed, but take a sequence of values of x_h so that $g(x) \to 1$.

Identification of the slope only without "identification at infinity"

If one is only interested in identifying the "slope" of g_f and not the intercept, one can avoid using an identification at infinity argument. That is, for any two points (x_f, x_0) and $(\tilde{x}_f, \tilde{x}_0)$, consider identifying the difference $g_f(x_f, x_0) - g_f(\tilde{x}_f, \tilde{x}_0)$. The key to identification is the existence of the exclusion restriction X_{hi}. For these two points, suppose we can find values x_h and $\tilde{x}_h$ so that

$$g(x_f, x_h, x_0) = g(\tilde{x}_f, \tilde{x}_h, \tilde{x}_0).$$

There may be many pairs of $(x_h, \tilde{x}_h)$ that satisfy this equality and we could choose any of them. Define $\tilde{x} \equiv (\tilde{x}_f, \tilde{x}_h, \tilde{x}_0)$. The key aspect of this is that since $g(x) = g(\tilde{x})$, and thus the probability of being a fisherman is the same given the two sets of points, then the bias terms are also the same: $\text{Med}(\varepsilon_{fi} \mid \varepsilon_i < g(x)) = \text{Med}(\varepsilon_{fi} \mid \varepsilon_i < g(\tilde{x}))$.

This allows us to write

$$\begin{aligned}
&\text{Med}(Y_i \mid X_i = x, J_i = f) - \text{Med}(Y_i \mid X_i = \tilde{x}, J_i = f) \\
&\quad = g_f(x_f, x_0) + \text{Med}(\varepsilon_{fi} \mid \varepsilon_i < g(x)) \\
&\qquad - [g_f(\tilde{x}_f, \tilde{x}_0) + \text{Med}(\varepsilon_{fi} \mid \varepsilon_i < g(\tilde{x}))] \\
&\quad = g_f(x_f, x_0) - g_f(\tilde{x}_f, \tilde{x}_0).
\end{aligned}$$

As long as we have sufficient variation in X_{hi} we can do this everywhere and identify g_f up to location.

Step 3: Identification of g_h

In terms of identifying g_h, the exclusion restriction that influences wages as a fisherman but not as a hunter (i.e. X_{fi}) will be crucial. Consider identifying $g_h(x_h, x_0)$ for any particular value (x_h, x_0). The key here is finding a value of x_f so that

$$\Pr(J_i = f \mid X_i = (x_f, x_h, x_0)) = 0.5. \tag{3.10}$$

Assumption 3.2 guarantees that we can do this. To see why Eq. (3.10) is useful, note that it must be that for this value of (x_f, x_h, x_0)

$$0.5 = \Pr\left(\varepsilon_{hi} - \varepsilon_{fi} \leq g_f(x_f, x_0) - g_h(x_h, x_0)\right). \tag{3.11}$$

But the fact that $\varepsilon_{hi} - \varepsilon_{fi}$ has median zero implies that

$$g_h(x_h, x_0) = g_f(x_f, x_0).$$

Since g_f is identified, g_h is identified from this expression.[8]

Again to relate this to the framework in Section 2.2 above, now $\Psi(\theta) = g_h$ and $\Psi(\Theta(p))$ is the set of functions g_h that are consistent with P. Above we showed that if $\Pr(J_i = f \mid X_i = x) = 0.5$, then $g_h(x_h, x_0) = g_f(x_f, x_0)$. Thus since we already showed that g_f is identified, g_h is the only element of $\Psi(\Theta(p))$.

Step 4: Identification of G

Next consider identification of G given g_f and g_h. We will show how to identify the joint distribution of $(\varepsilon_{fi}, \varepsilon_{hi})$ closely following the exposition of Heckman and Taber (2008). Note that from the data one can observe

$$\begin{aligned}
&\Pr(J_i = f, Y_{fi} < s \mid X_i = x) \\
&\quad = \Pr(g_h(x_h, x_0) + \varepsilon_{hi} \leq g_f(x_f, x_0) + \varepsilon_{fi}, g_f(x_f, x_0) + \varepsilon_{fi} \leq s) \\
&\quad = \Pr(\varepsilon_{hi} - \varepsilon_{fi} \leq g_f(x_f, x_0) - g_h(x_h, x_0), \varepsilon_{fi} \leq s - g_f(x_f, x_0)) \quad (3.12)
\end{aligned}$$

which is the cumulative distribution function of $(\varepsilon_{hi} - \varepsilon_{fi}, \varepsilon_{fi})$ evaluated at the point $(g_f(x_f, x_0) - g_h(x_h, x_0), s - g_f(x_f, x_0))$. By varying the point of evaluation one can identify the joint distribution of $(\varepsilon_{hi} - \varepsilon_{fi}, \varepsilon_{fi})$ from which one can derive the joint distribution of $(\varepsilon_{fi}, \varepsilon_{hi})$.

Finally in terms of the identification conditions in Section 2.2 above, now $\Psi(\theta) = G$ and $\Psi(\Theta(P))$ is the set of distributions G consistent with P. Since G is uniquely defined by the expression (3.12) and since everything else in this expression is identified, G is the only element of $\Psi(\Theta(P))$.

3.3. Relaxing independence between observables and unobservables

For expositional purposes we focus on the case in which the observables are independent of the unobservables, but relaxing these assumptions is easy to do. The simplest case is to allow for a general relationship between X_{0i} and $(\varepsilon_{fi}, \varepsilon_{hi})$. To see how easy this is, consider a case in which X_{0i} is just binary, for example denoting men and women. Independence seems like a very strong assumption in this case. For example, the distribution of unobserved preferences might be different for women and men, leading to different selection patterns. In order to allow for this, we could identify and estimate the Roy model separately for men and for women. Expanding from binary X_{0i} to finite support X_{0i} is trivial, and going beyond that to continuous X_{0i} is straightforward. Thus one can

[8] Note that Heckman and Honoré (1990) choose a different normalization. Rather than normalizing the median of $\varepsilon_{hi} - \varepsilon_{fi}$ to zero (which is convenient in the case in which Y_{hi} is not observed) they normalize the median of ε_{hi} to zero (which is more convenient in their case). Since this is just a normalization, it is innocuous. After identifying the model under our normalization we could go back to redefine the model in terms of theirs.

relax the independence assumption easily. But for expositional purposes we prefer our specification.

The distinction between X_{fi} and X_{0i} was not important in steps 1 and 2 of our discussion above. When one is only interested in the outcome equation $Y_{fi} = g_f(X_{fi}, X_{0i}) + \varepsilon_{fi}$, relaxing the independence assumption between X_{fi} and $(\varepsilon_{fi}, \varepsilon_{hi})$ can be done as well. However, in step 3 this distinction is important in identifying g_h and the independence assumption is not easy to relax.

If we allow for general dependence between X_{0i} and $(\varepsilon_{fi}, \varepsilon_{hi})$, the "identification at infinity" argument becomes more important as the argument about "Identification of the Slope Only without Identification at Infinity" no longer goes through. In that case the crucial feature of the model was that $\text{Med}(\varepsilon_{fi} \mid \varepsilon_i < g(x)) = \text{Med}(\varepsilon_{fi} \mid \varepsilon_i < g(\tilde{x}))$. However, without independence this is no longer generally true because $\text{Med}(\varepsilon_{fi} \mid X_i = x, J_i = f) = \text{Med}(\varepsilon_{fi} \mid X_{0i} = x_0, \varepsilon_i < g(x))$. Thus even if $g(x) = g(\tilde{x})$, when $x_0 \neq \tilde{x}_0$, in general $\text{Med}(\varepsilon_{fi} \mid X_{0i} = x_0, \varepsilon_i < g(x)) \neq \text{Med}(\varepsilon_{fi} \mid X_{0i} = \tilde{x}_0, \varepsilon_i < g(\tilde{x}))$.

3.4. The importance of exclusion restrictions

We now show that the model is not identified in general without an exclusion restriction.[9] Consider a simplified version of the model,

$$J_i = \begin{cases} f & \text{if } g(X_i) - \varepsilon_i \geq 0 \\ h & \text{otherwise} \end{cases}$$
$$Y_{fi} = g_f(X_i) + \varepsilon_{fi}$$

where ε_i is uniform (0,1) and $(\varepsilon_i, \varepsilon_{fi})$ is independent of X_i with distribution G and we use the location normalization $\text{Med}(\varepsilon_{fi} \mid X_i) = 0$. As in Section 3.2, we observe X_i, whether $J_i = f$ or h, and if $J_i = f$ then we observe Y_{fi}.

We can think about estimating the model from the median regression

$$\begin{aligned} \text{Med}[Y_{fi} | X_i = x] &= g_f(X_i) + \text{Med}[\varepsilon_{fi} | X_i = x] \\ &= g_f(X_i) + \text{Med}[\varepsilon_{fi} | g(X_i) > \varepsilon_i] \\ &= g_f(X_i) + h(g(X_i)). \end{aligned} \tag{3.13}$$

Under the assumption that $\text{Med}(\varepsilon_{fi} \mid X_i) = 0$ it must be the case that $h(1) = 0$, but this is our only restriction on h and g. Thus the model above has the same conditional

[9] An exception is Buera (2006), who allows for general functional forms and does not need an exclusion restriction. Assuming wages are observed in both sectors, and making stronger use of the independence assumption between the observables and the unobservables, he shows that the model can be identified without exclusion restrictions.

median as an alternative model

$$\text{Med}[Y_{fi}|X_i = x] = \widetilde{g}_f(X_i) + \widetilde{h}(g(X_i)) \tag{3.14}$$

where $\widetilde{g}_f(X_i) = g_f(X_i) + k(g(X_i))$ and $\widetilde{h}(g(X_i)) = h(g(X_i)) - k(g(X_i))$. Equations (3.13) and (3.14) are observationally equivalent. Without an exclusion restriction, it is impossible to tell if observed income from working varies with X_i because it varies with g_f or because it varies with the labor force participation rate and thus the extent of selection. Thus the models in Eqs (3.13) and (3.14) are not distinguishable using conditional medians.

To show the two models are indistinguishable using the full joint distribution of the data, consider an alternative data generating model with the same first stage, but now Y_{fi} is determined by

$$Y_{fi} = \widetilde{g}_f(X_i) + \widetilde{\varepsilon}_{fi}$$

where $\widetilde{\varepsilon}_{fi}$ is independent of X_i with $\text{Med}(\widetilde{\varepsilon}_{fi} \mid X_i) = 0$. Let $\widetilde{G}(\varepsilon_i, \widetilde{\varepsilon}_{fi})$ be the joint distribution of $(\varepsilon_i, \widetilde{\varepsilon}_{fi})$ in the alternative model. We will continue to assume that in the alternative model $\widetilde{g}_f(X_i) = g_f(X_i) + k(g(X_i))$. The question is whether the alternative model is able to generate the same data distribution.

In the true model

$$\begin{aligned}\Pr(\varepsilon_i \le g(x), Y_{fi} < y) &= \Pr(\varepsilon_i \le g(x), g_f(x) + \varepsilon_{fi} \le y) \\ &= G(g(x), y - g_f(x)).\end{aligned}$$

In the alternative model

$$\begin{aligned}\Pr(\varepsilon_i \le g(x), Y_{fi} < y) &= \Pr(\varepsilon_i \le g(x), \widetilde{g}_f(x) + \widetilde{\varepsilon}_{fi} \le y) \\ &= \widetilde{G}(g(x), y - \widetilde{g_f}(x)).\end{aligned}$$

Thus these two models generate exactly the same joint distribution of data and cannot be separately identified as long as we define $\widetilde{G}$ so that[10]

$$\begin{aligned}\widetilde{G}(g(x), y - \widetilde{g}_f(x)) &= G(g(x), y - g_f(x)) \\ &= G(g(x), y - \widetilde{g}_f(x) + k(g(x))).\end{aligned}$$

[10] One cannot do this with complete freedom as one needs $\widetilde{G}$ to be a legitimate cdf. That is, it must be nondecreasing in both of its arguments. However, there will typically be many examples of k for which $\widetilde{G}$ is a cdf and the model is not identified. For example, if k is a nondecreasing function, $\widetilde{G}$ will be a legitimate cdf.

4. THE GENERALIZED ROY MODEL

We next consider the "Generalized Roy Model" (as defined in e.g. (Heckman and Vytlacil, 2007a). The basic Roy model assumes that workers only care about their income. The Generalized Roy Model allows workers to care about non-pecuniary aspects of the job as well. Let U_{fi} and U_{hi} be the utility that individual i would receive from being a fisherman or a hunter respectively, where for $j \in \{f, h\}$,

$$U_{ji} = Y_{ji} + \varphi_j(Z_i, X_{0i}) + \nu_{ji}. \tag{4.1}$$

where $\varphi_j(Z_i, X_{0i})$ represents the non-pecuniary utility gain from observables Z_i and X_{0i}. The variable Z_i allows for the fact that there may be other variables that affect the taste for hunting versus fishing directly, but do not affect wages in either sector.[11] Note that we are imposing separability between Y_{ji} and φ_j. In general we can provide conditions in which the results presented here will go through if we relax this assumption, but we impose it for expositional simplicity. The occupation is now defined as

$$J_i = \begin{cases} f & \text{if } U_{fi} > U_{hi} \\ h & \text{if } U_{fi} \leq U_{hi}. \end{cases} \tag{4.2}$$

We continue to assume that

$$\begin{aligned} Y_{fi} &= g_f(X_{fi}, X_{0i}) + \varepsilon_{fi} \\ Y_{hi} &= g_h(X_{hi}, X_{0i}) + \varepsilon_{hi} \end{aligned} \tag{4.3}$$

$$Y_i = \begin{cases} Y_{fi} & \text{if } J_i = f \\ Y_{hi} & \text{if } J_i = h. \end{cases} \tag{4.4}$$

It will be useful to define a reduced form version of this model. Note that people fish when

$$\begin{aligned} 0 &< U_{fi} - U_{hi} \\ &= (Y_{fi} + \varphi_f(Z_i, X_{0i}) + \nu_{fi}) - (Y_{hi} + \varphi_h(Z_i, X_{0i}) + \nu_{hi}) \\ &= g_f(X_{fi}, X_{0i}) + \varphi_f(Z_i, X_{0i}) - g_h(X_{hi}, X_{0i}) - \varphi_h(Z_i, X_{0i}) \\ &\quad + \varepsilon_{fi} + \nu_{fi} - \varepsilon_{hi} - \nu_{hi}. \end{aligned}$$

In the previous section we described how the choice model can only be identified up to a monotonic transform and that assuming the error term is uniform is a convenient

[11] In principle some of the elements of Z_i may affect φ_f and others may affect φ_h, but this distinction will not be important here, so we use the most general notation.

normalization. We do the same thing here. Let F^* be the distribution function of $\varepsilon_{hi} + \nu_{hi} - \varepsilon_{fi} - \nu_{fi}$. Then we define

$$\nu_i \equiv F^*\left(\varepsilon_{hi} + \nu_{hi} - \varepsilon_{fi} - \nu_{fi}\right) \tag{4.5}$$

$$\varphi(Z_i, X_i) \equiv F^*(g_f(X_{fi}, X_{0i}) + \varphi_f(Z_i, X_{0i}) - g_h(X_{hi}, X_{0i}) - \varphi_h(Z_i, X_{0i})). \tag{4.6}$$

As above, this normalization is convenient because it is straightforward to show that

$$J_i = f \quad \text{when } \varphi(Z_i, X_i) > \nu_i$$

and that ν_i is uniformly distributed on the unit interval.

We assume that the econometrician can observe the occupations of the workers and the wages that they receive in their chosen occupations as well as (X_i, Z_i).

4.1. Identification

It turns out that the basic assumptions that allow us to identify the Roy model also allow us to identify the generalized Roy model.

We start with the reduced form model in which we need two more assumptions.

Assumption 4.1. $(\nu_i, \varepsilon_{fi}, \varepsilon_{hi})$ is continuously distributed and is independent of (Z_i, X_i). Furthermore, ν_i is distributed uniform on the unit interval and the medians of both ε_{fi} and ε_{hi} are zero.

Assumption 4.2. The support of $\varphi(Z_i, x)$ is $[0, 1]$ for all $x \in \text{supp}(X_i)$.

We also slightly extend the restrictions on the functions to include φ_f and φ_h.

Assumption 4.3. $(Z_i, X_i) = (Z_i, X_{fi}, X_{hi}, X_{0i})$ can be written as $(Z_i^c, Z_i^d, X_{fi}^c, X_{fi}^d, X_{hi}^c, X_{hi}^d, X_{0i}^c, X_{0i}^d)$ where the elements of $(Z_i^c, X_{fi}^c, X_{hi}^c, X_{0i}^c)$ are continuously distributed (no point has positive mass), and $(Z_i^d, X_{fi}^d, X_{hi}^d, X_{0i}^d)$ are distributed discretely (all support points have positive mass).

Assumption 4.4. For any $(z^d, x_f^d, x_h^d, x_0^d) \in \text{supp}(Z_i^d, X_{fi}^d, X_{hi}^d, X_{0i}^d)$, $g_f(x_f^c, x_f^d, x_0^c, x_0^d)$, $g_h(x_h^c, x_h^d, x_0^c, x_0^d)$, $\varphi_f(z^c, z^d, x_0^c, x_0^d)$ and $\varphi_h(z^c, z^d, x_0^c, x_0^d)$ are almost surely continuous across

$$(z^c, x^c) \in \text{supp}(Z_i^c, X_i^c \mid (Z_i^d, X_i^d) = (z^d, x^d)).$$

Theorem 4.1. *Under Assumptions 4.1–4.4, φ, g_f, g_h and the joint distribution of $(\nu_i, \varepsilon_{fi})$ and of $(\nu_i, \varepsilon_{hi})$ are identified from the joint distribution of (J_i, Y_i) on a set $\mathcal{X}^*$ that has measure 1 where (J_i, Y_i) are generated by model (4.1)–(4.4).*

(Proof in Appendix.)

The intuition for identification follows directly from the intuition given for the basic Roy model. We show this in 3 steps:

1. Identification of φ is like the "Step 1: identification of choice model" section. We can only identify φ up to a monotonic transformation for exactly the same reason given in that section. We impose the normalization that ν_i is uniform in Assumption 4.2. Given that assumption

$$\Pr(J_i = f \mid Z_i = z, X_i = x) = \varphi(z, x)$$

so identification of φ from $\Pr(J_i = f \mid Z_i = z, X_i = x)$ comes directly.

2. Identification of g_f and g_h are completely analogous to "Step 2: identification of g_f" in Section 3.2. That is

$$\begin{aligned}
&\lim_{\varphi(z,x)\to 1} \text{Med}(Y_i \mid Z_i = z, X_i = x, J_i = f) \\
&= g_f(x_f, x_0) + \lim_{\varphi(z,x)\to 1} \text{Med}(\varepsilon_{fi} \mid Z_i = z, X_i = x, J_i = f) \\
&= g_f(x_f, x_0) + \lim_{\varphi(z,x)\to 1} \text{Med}(\varepsilon_{fi} \mid \nu_i \le \varphi(z, x)) \\
&= g_f(x_f, x_0) + \text{Med}(\varepsilon_{fi}) \\
&= g_f(x_f, x_0).
\end{aligned}$$

The analogous argument works for g_h when we send $\varphi(z, x) \to 0$.

3. Identification of the joint distribution of $(\nu_i, \varepsilon_{fi})$ and of $(\nu_i, \varepsilon_{hi})$ are analogous to the "Step 4: identification of G" discussion in the Roy model. That is if we let G_{ν,ε_f} represent the joint distribution of $(\nu_i, \varepsilon_{fi})$ then

$$\begin{aligned}
&\Pr(J_i = f, Y_{fi} \le y \mid (Z_i, X_i) = (z, x)) \\
&\quad = \Pr(\nu_i \le \varphi(z, x), g_f(x_f, x_0) + \varepsilon_{fi} \le y) \\
&\quad = G_{\nu,\varepsilon_f}(\varphi(z, x), y - g_f(x_f, x_0)).
\end{aligned}$$

The analogous argument works for the joint distribution of $(\nu_i, \varepsilon_{hi})$.

Note that not all parameters are identified such as the non-pecuniary gain from fishing $\varphi_f - \varphi_h$. To identify the "structural" generalized Roy model we make two additional assumptions:

Assumption 4.5. The median of $\varepsilon_{hi} + \nu_{hi} - \varepsilon_{fi} - \nu_{fi}$ is zero.

Assumption 4.6. For any value of $(z, x_0) \in \text{supp}(Z_i, X_{0i})$, $g_f(X_{fi}, x_0) - g_h(X_{hi}, x_0)$ has full support (i.e. the whole real line).

Theorem 4.2. *Under Assumptions 4.1–4.6, $\varphi_f - \varphi_h$, the distribution of $(\varepsilon_{hi} + \nu_{hi} - \varepsilon_{fi} - \nu_{fi}, \varepsilon_{fi})$, and the distribution of $(\varepsilon_{hi} + \nu_{hi} - \varepsilon_{fi} - \nu_{fi}, \varepsilon_{hi})$ are identified.*

(Proof in Appendix.)

Note that Theorem 4.1 gives the joint distribution of $(\nu_i, \varepsilon_{fi})$ while Theorem 4.2 gives the joint distribution of $(\varepsilon_{hi} + \nu_{hi} - \varepsilon_{fi} - \nu_{fi}, \varepsilon_{fi})$. Since $\nu_i = F^*(\varepsilon_{hi} + \nu_{hi} - \varepsilon_{fi} - \nu_{fi})$, this really just amounts to saying that F^* is identified.

Furthermore, whereas g_f and g_h are identified in Theorem 4.1, $\varphi_f - \varphi_h$ is identified in Theorem 4.2. Recall $\varphi_f - \varphi_h$ is the added utility (measured in money) of being a fisherman relative to a hunter. The exclusion restrictions X_{fi} and X_{hi} help us identify this. These exclusion restrictions allow us to vary the pecuniary gains of the two sectors, holding preferences $\varphi_f - \varphi_h$ constant. Identification is analogous to the "Step 3: identification of g_h" in the standard Roy model. To see where identification comes from, for every (z, x_0) think about the following conditional median

$$\begin{aligned} 0.5 &= \Pr(J_i = f \mid Z_i = z, X_i = x) \\ &= \Pr(\varepsilon_{hi} + \nu_{hi} - \varepsilon_{fi} - \nu_{fi} \leq g_f(x_f, x_0) + \varphi_f(z, x_0) - g_h(x_h, x_0) - \varphi_h(z, x_0)). \end{aligned}$$

Since the median of $\varepsilon_{hi} + \nu_{hi} - \varepsilon_{fi} - \nu_{fi}$ is zero, this means that

$$g_f(x_f, x_0) + \varphi_f(z, x_0) - g_h(x_h, x_0) - \varphi_h(z, x_0) = 0,$$

and thus

$$\varphi_f(z, x_0) - \varphi_h(z, x_0) = g_h(x_h, x_0) - g_f(x_f, x_0).$$

Because g_f and g_h is identified, $\varphi_f - \varphi_h$ is identified also. The argument above shows that we do not need both X_{fi} and X_{hi}, we only need X_{fi} or X_{hi}.

Suppose there is no variable that affects earnings in one sector but not preferences (X_{fi} or X_{hi}). An alternative way to identify $\varphi_f - \varphi_h$ is to use a cost measured in dollars. Consider the linear version of the model with normal errors and without exclusion restrictions (X_{hi}, X_{fi}) so that

$$\begin{aligned} g_h(x_0) &= x_{0i}'\gamma_h \\ g_f(x_0) &= x_{0i}'\gamma_f \\ \varphi_f(z, x_0) - \varphi_h(z, x_0) &= x_0'\beta_0 + z'\beta_z. \end{aligned}$$

The reduced form probit is:

$$\Pr(J_i = f \mid Z_i = z, X_i = x) = \Phi\left(x_{0i}' \frac{\gamma_f - \gamma_h + \beta_0}{\sigma} + z_i' \frac{\beta_z}{\sigma}\right)$$

where σ is the standard deviation of $\varepsilon_{hi} + \nu_{hi} - \varepsilon_{fi} - \nu_{fi}$. Theorem 4.1 above establishes that the functions g_f and g_h (i.e., γ_f and γ_h) as well as the variance of ε_{hi} and ε_{fi} are identified. We still need to identify β_0, β_z and σ. Thus we are able to identify

$$\frac{\gamma_f - \gamma_h + \beta_0}{\sigma} \quad \text{and} \quad \frac{\beta_z}{\sigma}.$$

If β_0 and β_z are scalars we still have three parameters $(\beta_0, \beta_z, \sigma)$ and two restrictions $(\frac{\gamma_f - \gamma_h + \beta_0}{\sigma}, \frac{\beta_z}{\sigma})$. If they are not scalars, we still have one more parameter than restriction. However suppose that one of the exclusion restrictions represents a cost variable that is measured in the same units as $Y_{fi} - Y_{hi}$. For example in a schooling case suppose that Y_{fi} represents the present value of earnings as a college graduate, Y_{hi} represents the present value of high school graduate as a college graduate, and the exclusion restriction, Z_i, represents the present value of college tuition. In this case $\beta_z = -1$ the coefficient on Z_i is $-1/\sigma$, so σ is identified. Given σ it is very easy to show that the rest of the parameters are identified as well. Heckman et al. (1998) provide an example of this argument using tuition as in the style above. In Section 7.3 we discuss Heckman and Navarro (2007) who use this approach as well.

4.2. Lack of identification of the joint distribution of $(\varepsilon_{fi}, \varepsilon_{hi})$

In pointing out what is identified in the model it is also important to point out what is not identified. Most importantly in the generalized Roy model we were able to identify the joint distribution between the error terms in the selection equation and each of the outcomes, but not the joint distribution of the variables in the outcome equation. In particular the joint distribution between the error terms $(\varepsilon_{fi}, \varepsilon_{hi})$ is not identified. Even strong functional form assumptions will not solve this problem. Fir example, it is easy to show that in the joint normal model the covariance of $(\varepsilon_{fi}, \varepsilon_{hi})$ is not identified.

4.3. Are functional forms innocuous? Evidence from Catholic schools

As the theorems above make clear, nonparametric identification requires exclusion restrictions. However, completely parametric models typically do not require exclusion restrictions. In specific empirical examples, identification could primarily be coming from the exclusion restriction or identification could be coming primarily from the functional form assumptions (or some combination between the two). When researchers use exclusion restrictions in data, it is important to be careful about which assumptions are important.

We describe one example from Altonji et al. (2005b). Based on Evans and Schwab (1995), Neal (1997), and Neal and Grogger (2000) they consider a bivariate probit model of Catholic schooling and college attendance.

$$CH_i = 1(X_i'\beta + \lambda Z_i + u_i > 0) \tag{4.7}$$

$$Y_i = 1(\alpha CH_i + X_i'\gamma + \varepsilon_i > 0), \tag{4.8}$$

where $1(\cdot)$ is the indicator function taking the value one if its argument is true and zero otherwise, CH_i is a dummy variable indicating attendance at a Catholic school, and Y_i is a dummy variable indicating college attendance. Identification of the effect of Catholic schooling on college attendance (or high school graduation) is the primary focus of these studies. The question at hand is in practice whether the assumed functional forms for u_i and ε_i are important for identifying the α coefficient and thus the effect of Catholic schools on college attendance.

The model in Eqs (4.7) and (4.8) is a minor extension of the generalized Roy model. The first key difference is that the outcome variable in Eq. (4.8) is binary (attend college or not), whereas in the case of the Generalized Roy model the outcomes were continuous (earnings in either sector). The second key difference is that the outcome equation for Catholic versus Non-Catholic school only differs in the intercept (α). The error term (ε_i) and the slope coefficients (γ) are restricted to be the same. Nevertheless, the machinery to prove non-parametric identification of the Generalized Roy model can be applied to this framework.[12]

Using data from the National Longitudinal Survey of 1972, Altonji et al. (2005b) consider an array of instruments and different specifications for Eqs (4.7) and (4.8). In Table 1 we present a subset of their results. We show four different models. The "Single Equation Model" gives results in which selection into Catholic school is not accounted for. The first column gives results from a probit model (with point estimates, standard errors, and marginal effects). The second column give results from a Linear Probability model. Next we present the estimates of α from a Bivariate Probit models with alternative exclusion restrictions. The final row presents the results with no exclusion restrictions. Finally we also present results from an instrumental variable linear probability model with the same set of exclusion restrictions.

One can see that the marginal effect from the single equation probit is very similar to the OLS estimate. It indicates that college attendance rates are approximately 23.9 percentage points higher for Catholic high school graduates than for public high school graduates. The rest of the table presents results from three bivariate probit models and two instrumental variables models using alternative exclusion restrictions. The problem is clearest when the interaction between the student coming from a Catholic school and distance to the nearest Catholic school is used as an instrument. The 2SLS gives nonsensical results: a coefficient of 2.572 with an enormous standard error. This indicates that the instrument has little power. However, the bivariate probit result is more reasonable. It suggests that the true marginal causal effect is around 0.478 and the point

[12] Following Matzkin (1992), we need a monotonic normalization on the outcome model (such as assuming the error term is uniform). Once we have done this, proving identification of this model is almost identical to the generalized Roy model and is easily done with an exclusion restriction with sufficient support.

Table 1 Estimated effects of Catholic schools on college attendance from linear and nonlinear models.

Single equation models		
	Probit	OLS
	0.239	0.239
	[0.640]	
	(0.198)	(0.070)
Two equation models		
Excluded variable	Bivariate probit	2SLS
Catholic	0.285	−0.093
	[0.761]	
	(0.543)	(0.324)
Catholic × Distance	0.478	2.572
	[1.333]	
	(0.516)	(2.442)
None	0.446	
	[1.224]	
	(0.542)	

Urban Non-Whites from NLS-72.
The first set of results come from simple probits and from OLS.
The further results come from Bivariate Probits and from two stage least squares.
We present the marginal effect of Catholic high school attendance on college attendance.
[Point Estimate from Probit in Brackets.]
(Standard Errors in Parentheses.)
Source: Altonji et al. (2005b).

estimate is statistically significant. This seems inconsistent with the 2SLS results which indicated that this exclusion restriction had very little power. However it is clear what is going on when we compare this result to the model at the bottom of the table without an exclusion restriction. The estimate is very similar with a similar standard error. The linearity and normality assumptions drive the results.

The case in which Catholic religion by itself is used as an instrument is less problematic. The IV result suggests a strong amount of positive selection but still yields a large standard error. The bivariate probit model suggests a marginal effect that is a bit larger than the OLS effect. However, note that the standard errors for the model with and without an exclusion restriction are quite similar, which seems inconsistent with the idea that the exclusion restriction is providing a lot of identifying information. Further note that the IV result suggests a strong positive selection bias while the bivariate probit without exclusion restrictions suggests a strong negative bias. The bivariate probit in which Catholic is excluded is somewhere between the two. This suggests that both functional form and exclusion restrictions are important in this case. We should emphasize the "suggests" part of this sentence as none of this is a formal test. It does,

however, make one wonder how much trust to put in the bivariate probit results by themselves.

Another paper documenting the importance of functional form assumptions is Das et al. (2003), who estimate the return to education for young Australian women. They estimate equations for years of education, the probability of working, and wages. When estimating the wage equation they address both the endogeneity of years of education and also selection caused because we only observe wages for workers. They allow for flexibility in the returns to education (where the return depends on years of education) and also in the distribution of the residuals. They find that when they assume normality of the error terms, the return to education is approximately 12%, regardless of years of education. However, once they allow for more flexible functional forms for the error terms, they find that the returns to education decline sharply with years of education. For example, they find that at 10 years of education, the return to education is over 15%. However, at 14 years, the return to education is only about 5%.

5. TREATMENT EFFECTS

There is a very large literature on the estimation of treatment effects. For more complete summaries see Heckman and Robb (1986), Heckman et al. (1999), Heckman and Vytlacil (2007a,b), Abbring and Heckman (2007), or Imbens and Wooldridge (2009).[13] DiNardo and Lee (2011) provide a discussion that is complementary to ours. Our goal in this section is not to survey the whole literature but provide a brief summary and to put it into the context of identification of the Generalized Roy Model.

The goal of this literature is to estimate the value of receiving a treatment defined as:

$$\pi_i = Y_{fi} - Y_{hi}. \tag{5.1}$$

In the context of the Roy model, π_i is the income gain from moving from hunting to fishing. This income gain potentially varies across individuals in the population. Thus for people who choose to be fishermen, π_i is positive and for people who choose to be hunters, π_i is negative.

Estimation of treatment effects is of great interest in many literatures. The term "treatment effect"makes the most sense in the context of the medical literature. Choice f could represent taking a medical treatment (such as an experimental drug) while h could represent no treatment. In that case Y_{fi} and Y_{hi} would represent some measure of health status for individual i with and without the treatment. Thus the treatment effect π_i is the effect of the drug on the health outcome for individual i.

[13] There is also a substantial literature on the tradeoffs between different empirical approaches. Key papers include Leamer (1983), Heckman (1979, 1999, 2000), Angrist and Imbens (1999), Rosenzweig and Wolpin (2000), Deaton (2009), Heckman and Urzúa (2010), Imbens (2009), Angrist and Pischke (2010) and Sims (2010).

The classic example in labor economics is job training. In that case, Y_{fi} would represent a labor market outcome for individuals who received training and Y_{hi} would represents the outcome in the absence of training.

In both the case of drug treatment and job training, empirical researchers have exploited randomized trials. Medical patients are often randomly assigned either a treatment or a placebo (i.e., a sugar pill that should have no effect on health). Likewise, many job training programs are randomly assigned. For example, in the case of the Job Training Partnership Act, a large number of unemployed individuals applied for job training (see e.g. Bloom et al., 1997). Of those who applied for training, some were assigned training and some were assigned no training.

Because assignment is random and affects the level of treatment, one can treat assignment as an exclusion restriction that is correlated with treatment (i.e., the probability that $J_i = f$) but is uncorrelated with preferences or ability because it is random. In this sense, random assignment solves the selection problem that is the focus of the Roy model. As we show below, exogenous variation provided by experiments allows the researcher to cleanly identify some properties of the distribution of Y_{fi} and Y_{hi} under relatively weak assumptions. Furthermore, the methods for estimating these objects are simple, which adds to their appeal.

The treatment effect framework is also widely used for evaluating quasi-experimental data as well. By quasi-experimental data, we mean data that are not experimental, but exploit variation that is "almost as good as" random assignment.

5.1. Treatment effects and the generalized Roy model

Within the context of the generalized Roy model note that in general

$$\pi_i = g_f(X_{fi}, X_{0i}) - g_h(X_{hi}, X_{0i}) + \varepsilon_{fi} - \varepsilon_{hi}.$$

An important special case of the treatment effect defined in Eq. (5.1) is when

$$g_f(X_{fi}, X_{0i}) = g_h(X_{hi}, X_{0i}) + \pi_0 \tag{5.2}$$

$$\varepsilon_{fi} = \varepsilon_{hi}. \tag{5.3}$$

In this case, the treatment effect $\pi_i = Y_{fi} - Y_{hi} = \pi_0$ is a constant across individuals. Identification of this parameter is relatively straightforward. However, there is a substantial literature that studies identification of heterogeneous treatment effects. As we point out above, treatment effects are positive for some people and negative for others in the context of the Roy model. Furthermore, there is ample empirical evidence that the returns to job training are not constant, but instead vary across the population (Heckman et al., 1999).

In Section 4.2 we explain why the joint distribution of $(\varepsilon_{fi}, \varepsilon_{hi})$ is not identified. This means that the distribution of π_i is not identified and even relatively simple summary statistics like the median of this distribution is not identified in general. The key problem is that even when assignment is random, we do not observe the same people in both occupations.

Since the full generalized Roy model is complicated, hard to describe, and very demanding in terms of data, researchers often focus on a summary statistic to summarize the result. The most common in this literature is the Average Treatment Effect (ATE) defined as

$$\begin{aligned} \text{ATE} &\equiv E(\pi_i) \\ &= E(Y_{fi}) - E(Y_{hi}). \end{aligned}$$

From Theorem 4.1 we know that (under the assumptions of that theorem) the distribution of Y_{fi} and Y_{hi} are identified. Thus, their expected values are also identified under the one additional assumption that these expected values exist.

Assumption 5.1. The expected values of Y_{fi} and Y_{hi} are finite.

Theorem 5.1. *Under the assumptions of Theorem 4.1 and Assumption 5.1, the Average Treatment effect is identified.*

(Proof in Appendix.)

To see where identification of this object comes from, abstract from X_i so that the only observable is Z_i, which affects the non-pecuniary gain in utility from occupation across occupations. With experimental data, Z_i could be randomly generated assignments to occupation. Notice that

$$\begin{aligned} &\lim_{\varphi(z)\to 1} E(Y_{fi} \mid Z_i = z, J_i = f) - \lim_{\varphi(z)\to 0} E(Y_{hi} \mid Z_i = z, J_i = h) \\ &= \lim_{\varphi(z)\to 1} E(Y_{fi} \mid \nu_i \le \varphi(z)) - \lim_{\varphi(z)\to 0} E(Y_{hi} \mid \nu_i > \varphi(z)) \\ &= E(Y_{fi}) - E(Y_{hi}). \end{aligned}$$

Thus the exclusion restriction is the key to identification. Note also that we need groups of individuals where $\varphi(Z_i) \approx 1$ (who are always fishermen) and $\varphi(Z_i) \approx 0$ (who are always hunters); thus "identification at infinity" is essential as well. For the reasons discussed in the nonparametric Roy model above, if $\varphi(Z_i)$ were never higher than some $\varphi(z^u) < 1$ then $E(Y_{fi})$ would not be identified. Similarly if $\varphi(Z_i)$ were never lower than some $\varphi(z^\ell) > 0$, then $E(Y_{hi})$ would not be identified.

While one could directly estimate the ATE using "identification at infinity", as described above, this is not the common practice and not something we would advocate.

The standard approach would be to estimate the full Generalized Roy Model and then use it to simulate the various treatment effects. This is often done using a completely parametric approach as in, for example, the classic paper by Willis and Rosen (1979). However, there are quite a few nonparametric alternatives as well, including construction of the Marginal Treatment effects as discussed in Sections 5.3 and 5.4 below.

As it turns out, even with experimental data, it is rarely the case that $\varphi(Z_i)$ is identically one or zero with positive probability. In the case of medicine, some people assigned the treatment do not take the treatment. In the training example, many people who are offered subsidized training decide not to undergo the training. Thus, when compliance with assignment is less than 100%, we cannot recover the ATE. In Section 5.2 we discuss more precisely what we do recover when there is less than 100% compliance.

It is also instructive to relate the ATE to instrumental variables estimation. Let Y_i be the outcome of interest

$$Y_i = \begin{cases} Y_{fi} & \text{if } J_i = f \\ Y_{hi} & \text{if } J_i = h, \end{cases}$$

and let D_{fi} be a dummy variable indicating whether $J_i = f$. Consider estimating the model

$$Y_i = \beta_0 + \beta_1 D_{fi} + u_i \tag{5.4}$$

using instrumental variables with Z_i as an instrument for D_{fi}. Assume that Z_i is correlated with D_{fi} but not with Y_{fi} or Y_{hi}. Consider first the constant treatment effect model described in Eqs (5.2) and (5.3) so that $\pi_i = \pi_0$ for everyone in the population. In that case

$$\begin{aligned} Y_i &= Y_{fi} D_{fi} + Y_{hi}(1 - D_{fi}) \\ &= Y_{hi} + D_{fi}(Y_{fi} - Y_{hi}) \\ &= Y_{hi} + D_{fi}\pi_0. \end{aligned}$$

Then two stage least squares on the model above yields

$$\begin{aligned} \operatorname{plim} \widehat{\beta}_1 &= \frac{\operatorname{cov}(Z_i, Y_i)}{\operatorname{cov}(Z_i, D_{fi})} \\ &= \frac{\operatorname{cov}(Z_i, Y_{hi} + D_{fi}\pi_0)}{\operatorname{cov}(Z_i, D_{fi})} \\ &= \frac{\operatorname{cov}(Z_i, Y_{hi})}{\operatorname{cov}(Z_i, D_{fi})} + \frac{\operatorname{cov}(Z_i, \pi_0 D_{fi})}{\operatorname{cov}(Z_i, D_{fi})} \\ &= \pi_0. \end{aligned}$$

Thus in the constant treatment effect model, instrumental variables provide a consistent estimate of the treatment effect. However, this result does not carry over to heterogeneous treatment effects or the average treatment effects as Heckman (1997) shows. Following the expression above we get

$$\begin{aligned}\operatorname{plim}\widehat{\beta}_1 &= \frac{\operatorname{cov}(Z_i, Y_{hi} + D_{fi}\pi_i)}{\operatorname{cov}(Z_i, D_{fi})} \\ &= \frac{\operatorname{cov}(Z_i, D_{fi}\pi_i)}{\operatorname{cov}(Z_i, D_{fi})} \\ &\neq \text{ATE} \end{aligned} \tag{5.5}$$

in general. In Sections 5.2 and 5.3 below, we describe what instrumental variables identify.

In practice there are two potential problems with the assumptions behind Theorem 5.1 above

- The researcher may not have a valid exclusion restriction. We discuss some of the options for this case in Sections 5.5–5.7.
- Even if they do, the variable may not have full support. By this we mean that the instrumental variable Z_i may not vary enough, so that for some observed values of Z_i everyone is always a fisherman and for other observed values of Z_i everyone is always a hunter. We discuss what can be identified using exclusion restrictions with limited support in Sections 5.2–5.4 and 5.6.

We discuss a number of different approaches, some of which assume an exclusion restriction but relax the support conditions and others that do not require exclusion restrictions.

5.2. Local average treatment effects

Imbens and Angrist (1994) and Angrist et al. (1996) consider identification when the support of Z_i takes on a finite number of points. They show that when varying the instrument over this range, they can identify what they call a Local Average Treatment Effect. Furthermore, they show how instrumental variables can be used to estimate it. It is again easiest to think about this problem after abstracting from X_i, as it is straightforward to condition on these variables (see Imbens and Angrist, 1994, for details). For simplicity's sake, consider the case in which the instrument Z_i is binary and takes on the values $\{0, 1\}$. In many cases not only is the instrument discrete, but it is also binary. For example, in randomized medical trials, $Z_i = 1$ represents assignment to treatment, whereas $Z_i = 0$ represents assignment to the placebo. In job training programs, $Z_i = 1$ represents assignment to the training program, whereas $Z_i = 0$ represents no assigned training.

It is important to point out that not all patients assigned treatment actually receive the treatment. Thus $J_i = f$ if the patient actually takes the drug and $J_i = h$ if the individual does not take the drug. Likewise, not all individuals who are assigned training actually receive the training, so $J_i = f$ if the individual goes to training and $J_i = h$ if she does not. The literature on Local Average Treatment Effects handles this case as well as many others. However, we do require that the instrument of assignment has power: $\Pr(J_i = f \mid Z_i = 1) \neq \Pr(J_i = f \mid Z_i = 0)$. Without loss of generality we will assume that $\Pr(J_i = f \mid Z_i = 1) > \Pr(J_i = f \mid Z_i = 0)$.

Using the reduced form version of the generalized Roy model the choice problem is

$$J_i = f \quad \text{if } \varphi(Z_i) > \nu_i \tag{5.6}$$

where ν_i is uniformly distributed.

The following six objects can be learned directly from the data:

$$\begin{aligned}
&\Pr(J_i = f|Z_i = 0) = \Pr(\nu_i \leq \varphi(0)) \\
&\Pr(J_i = f|Z_i = 1) = \Pr(\nu_i \leq \varphi(1)) \\
&E(Y_{fi} \mid Z_i = 0, J_i = f) = E(Y_{fi} \mid \nu_i \leq \varphi(0)) \\
&E(Y_{hi} \mid Z_i = 0, J_i = h) = E(Y_{hi} \mid \nu_i > \varphi(0)) \\
&E(Y_{fi} \mid Z_i = 1, J_i = f) = E(Y_{fi} \mid \nu_i \leq \varphi(1)) \\
&E(Y_{hi} \mid Z_i = 1, J_i = h) = E(Y_{hi} \mid \nu_i > \varphi(1)).
\end{aligned}$$

The above equations show that our earlier assumption that $\Pr(J_i = f|Z_i = 1) > \Pr(J_i = f|Z_i = 0)$ implies $\Pr(\nu_i \leq \varphi(1)) > \Pr(\nu_i \leq \varphi(0))$. This, combined with the structure embedded in Eq. (5.6) means that

$$\Pr(\nu_i \leq \varphi(1)|\nu_i \leq \varphi(0)) = 1, \tag{5.7}$$

so then an individual who is a fisherman when $Z_i = 0$ is also a fisherman when $Z_i = 1$. Similar reasoning implies $\Pr(\nu_i \leq \varphi(1)|\varphi(0) < \nu_i \leq \varphi(1)) = 1$. Using this and Bayes rule yields

$$\begin{aligned}
\Pr(\nu_i \leq \varphi(0) \mid \nu_i \leq \varphi(1)) &= \frac{\Pr(\nu_i \leq \varphi(1) \mid \nu_i \leq \varphi(0)) \Pr(\nu_i \leq \varphi(0))}{\Pr(\nu_i \leq \varphi(1))} \\
&= \frac{\Pr(\nu_i \leq \varphi(0))}{\Pr(\nu_i \leq \varphi(1))},
\end{aligned} \tag{5.8}$$

$$\begin{aligned}
&\Pr(\varphi(0) < \nu_i \leq \varphi(1) \mid \nu_i \leq \varphi(1)) \\
&\quad = \frac{\Pr(\nu_i \leq \varphi(1) \mid \varphi(0) < \nu_i \leq \varphi(1)) \Pr(\varphi(0) < \nu_i \leq \varphi(1))}{\Pr(\nu_i \leq \varphi(1))} \\
&\quad = \frac{\Pr(\varphi(0) < \nu_i \leq \varphi(1))}{\Pr(\nu_i \leq \varphi(1))}.
\end{aligned} \tag{5.9}$$

Using the fact that $\Pr(\nu_i \leq \varphi(1)) = \Pr(\nu_i \leq \varphi(0)) + \Pr(\varphi(0) < \nu_i \leq \varphi(1))$, one can show that

$$\begin{aligned}
E(Y_{fi} \mid \nu_i \leq \varphi(1)) &= E(Y_{fi} \mid \nu_i \leq \varphi(0)) \Pr(\nu_i \leq \varphi(0) \mid \nu_i \leq \varphi(1)) \\
&\quad + E(Y_{fi} \mid \varphi(0) < \nu_i \leq \varphi(1)) \Pr(\varphi(0) < \nu_i \leq \varphi(1) \mid \nu_i \leq \varphi(1)).
\end{aligned} \tag{5.10}$$

Combining Eq. (5.10) with Eqs (5.8) and (5.9) yields

$$\begin{aligned}
E(Y_{fi} \mid \nu_i \leq \varphi(1)) &= \frac{E(Y_{fi} \mid \nu_i \leq \varphi(0)) \Pr(\nu_i \leq \varphi(0))}{\Pr(\nu_i \leq \varphi(1))} \\
&\quad + \frac{E(Y_{fi} \mid \varphi(0) < \nu_i \leq \varphi(1)) \Pr(\varphi(0) < \nu_i \leq \varphi(1))}{\Pr(\nu_i \leq \varphi(1))}.
\end{aligned} \tag{5.11}$$

Rearranging Eq. (5.11) shows that we can identify

$$\begin{aligned}
&E(Y_{fi} \mid \varphi(0) \leq \nu_i < \varphi(1)) \\
&= \frac{E(Y_{fi} \mid Z_i = 1, J_i = f) \Pr(J_i = f \mid Z_i = 1) - E(Y_{fi} \mid Z_i = 0, J_i = f) \Pr(J_i = f \mid Z_i = 0)}{\Pr(J_i = f \mid Z_i = 1) - \Pr(J_i = f \mid Z_i = 0)}
\end{aligned} \tag{5.12}$$

since everything on the right hand side is directly identified from the data.

Using the analogous argument one can show that

$$\begin{aligned}
&E(Y_{hi} \mid \varphi(0) \leq \nu_i < \varphi(1)) \\
&= \frac{E(Y_{hi} \mid Z_i = 0, J_i = h) \Pr(J_i = h \mid Z_i = 0) - E(Y_{hi} \mid Z_i = 1, J_i = h) \Pr(J_i = h \mid Z_i = 1)}{\Pr(J_i = f \mid Z_i = 1) - \Pr(J_i = f \mid Z_i = 0)}
\end{aligned}$$

is identified. But this means that we can identify

$$E(\pi_i \mid \varphi(0) \leq \nu_i < \varphi(1)) = E(Y_{fi} - Y_{hi} \mid \varphi(0) \leq \nu_i < \varphi(1)) \tag{5.13}$$

which Imbens and Angrist (1994) define as the Local Average Treatment Effect. This is the average treatment effect for that group of individuals who would alter their treatment status if their value of Z_i changed. Given the variation in Z_i, this is the only group for whom we can identify a treatment effect. Any individual in the data with $\nu_i > \varphi(1)$

would never choose $J_i = f$, so the data are silent about $E(Y_{fi} \mid \nu_i > \varphi(1))$. Similarly the data is silent about $E(Y_{hi} \mid \nu_i \leq \varphi(0))$.

Imbens and Angrist (1994) also show that the standard linear Instrumental Variables estimator yield consistent estimates of Local Average Treatment Effects. Consider the instrumental variables estimator of Eq. (5.4)

$$Y_i = \beta_0 + \beta_1 D_{fi} + u_i.$$

In Eq. (5.5) we showed that

$$\begin{aligned}\widehat{\beta}_1 &\xrightarrow{p} \frac{\text{cov}(Z_i, D_{fi}\pi_i)}{\text{cov}(Z_i, D_{fi})} \\ &= \frac{E(\pi_i D_{fi} Z_i) - E\left(\pi_i D_{fi}\right) E\left(Z_i\right)}{E(D_{fi} Z_i) - E\left(D_{fi}\right) E\left(Z_i\right)}.\end{aligned}$$

Let P_z denote the probability that $Z_i = 1$. The numerator of the above expression is

$$\begin{aligned}&E(\pi_i D_{fi} Z_i) - E(\pi_i D_{fi}) E\left(Z_i\right) \\ &\quad = P_z E(\pi_i D_{fi} \mid Z_i = 1) - E\left(\pi_i D_{fi}\right) P_z \\ &\quad = P_z E(\pi_i D_{fi} \mid Z_i = 1) \\ &\qquad - \left[P_z E(\pi_i D_{fi} \mid Z_i = 1) + (1 - P_z) E(\pi_i, D_{fi} \mid Z_i = 0)\right] P_z \\ &\quad = P_z(1 - P_z)\left[E(\pi_i D_{fi} \mid Z_i = 1) - E(\pi_i D_{fi} \mid Z_i = 0)\right] \\ &\quad = P_z(1 - P_z) E(\pi_i \mid \varphi(0) < \nu_i \leq \varphi(1)) \Pr(\varphi(0) < \nu_i \leq \varphi(1))\end{aligned}$$

where the key simplification comes from the fact that

$$\begin{aligned}E(\pi_i D_{fi} \mid Z_i = 1) &= E\left(\pi_i 1\left(\nu_i \leq \varphi(1)\right)\right) \\ &= E\left(\pi_i\left[1\left(\nu_i \leq \varphi(0)\right) + 1\left(\varphi(0) < \nu_i \leq \varphi(1)\right)\right]\right) \\ &= E(\pi_i D_{fi} \mid Z_i = 0) \\ &\quad + E(\pi_i \mid \varphi(0) < \nu_i \leq \varphi(1)) \Pr(\varphi(0) < \nu_i \leq \varphi(1)).\end{aligned}$$

Next consider the denominator

$$\begin{aligned}&E(D_{fi} Z_i) - E(D_{fi}) E\left(Z_i\right) \\ &\quad = P_z E(D_{fi} \mid Z_i = 1) - E(D_{fi}) P_z \\ &\quad = P_z E(D_{fi} \mid Z_i = 1) - \left[P_z E(D_{fi} \mid Z_i = 1) + (1 - P_z) E(D_{fi} \mid Z_i = 0)\right] P_z\end{aligned}$$

$$= P_z(1 - P_z)\left[E(D_{fi} \mid Z_i = 1) - E(D_{fi} \mid Z_i = 0)\right]$$
$$= P_z(1 - P_z)\Pr(\varphi(0) < \nu_i \leq \varphi(1)).$$

Thus

$$\widehat{\beta}_1 \xrightarrow{p} \frac{E(\pi_i D_{fi} Z_i) - E(\pi_i D_{fi})E(Z_i)}{E(D_{fi} Z_i) - E(D_{fi})E(Z_i)}$$
$$= \frac{P_z(1 - P_z)E(\pi_i \mid \varphi(0) < \nu_i \leq \varphi(1))\Pr(\varphi(0) < \nu_i \leq \varphi(1))}{P_z(1 - P_z)\Pr(\varphi(0) < \nu_i \leq \varphi(1))}$$
$$= E(\pi_i \mid \varphi(0) < \nu_i \leq \varphi(1)).$$

Imbens and Angrist never explicitly use the generalized Roy model or the latent index framework. Instead, they write their problem only in terms of the choice probabilities. However, in order to do this they must make one additional assumption. Specifically, they assume that if $J_i = f$ when $Z_i = 0$ then $J_i = f$ when $Z_i = 1$. Thus changing $Z_i = 0$ to $Z_i = 1$ never causes some people to switch from fishing to hunting. It only causes people to switch from hunting to fishing. They refer to this as a monotonicity assumption. Vytlacil (2002) points out that this is implied by the latent index model when the index $\varphi(Z_i)$ is separable from ν_i, as we assumed in Eq. (5.6). As is implied by Eq. (5.7), increasing the index $\varphi(Z_i)$ will cause some people to switch from hunting to fishing, but not the reverse.[14]

Throughout, we use the latent index framework that is embedded in the Generalized Roy model, for three reasons. First, we can appeal to the identification results of the Generalized Roy model. Second, the latent index can be interpreted as the added utility from making a decision. Thus we can use the estimated model for welfare analysis. Third, placing the choice in an optimizing framework allows us to test the restrictions on choice that come from the theory of optimization.

As we have pointed out, not everyone offered training actually takes the training. For example, in the case of the JTPA, only 60% of those offered the training actually received it (Bloom et al., 1997). Presumably, those who took the training are those who stood the most to gain from the training. For example, the reason that many people do not take training is that they receive a job offer before training begins. For these people, the training may have been of relatively little value. Furthermore, 2% of those who applied for and were not assigned training program wind up receiving the training (Bloom et al., 1997). Angrist et al. (1996) refer to those who were assigned training, but did not take the training as *never-takers*. Those who receive the training whether or not

[14] However, he points out that the non-separable model $D_{fi} = 1(f(Z_i, \nu_i) > 0)$ does not necessarily give rise to monotonicity. All other differences between the latent variable framework and the LATE framework are extremely technical and minor.

they are assigned are *always-takers*. Those who receive the training only when assigned the training are *compliers*. In terms of the latent index framework, the never-takers are those for whom ($\nu_i \geq \varphi(1)$), the compliers are those for whom ($\varphi(0) \leq \nu_i < \varphi(1)$), and the always-takers are those for whom ($\nu_i < \varphi(0)$).

The monotonicity assumption embedded in the latent index framework rules out the existence of a final group: the *defiers*. In the context of training, this would be an individual who receives training when not assigned training but would not receive training when assigned. At least in the context of training programs (and many other contexts) it seems safe to assume that there are no defiers.

5.3. Marginal treatment effects

Heckman and Vytlacil (1999, 2001, 2005, 2007b) develop a framework that is useful for constructing many types of treatment effects. They focus on the marginal treatment effect (MTE) defined in our context as

$$\Delta^{\text{MTE}}(x, \nu) \equiv E(\pi_i \mid X_i = x, \nu_i = \nu).$$

They show formally how to identify this object. We present their methodology using our notation.

Note that if we allow for regressors X_i, let the exclusion restriction Z_i to take on values beyond zero and one, then if (z^ℓ, x) and (z^h, x) are in the support of the data, then Eq. (5.12) can be rewritten as

$$\begin{aligned} &E(Y_{fi} \mid \varphi(z^\ell, x) \leq \nu_i < \varphi(z^h, x), X_i = x) \\ &= \frac{E(Y_{fi} \mid (Z_i, X_i) = (z^h, x), J_i = f) \Pr(J_i = f \mid (Z_i, X_i) = (z^h, x))}{\Pr(J_i = f \mid (Z_i, X_i) = (z^h, x)) - \Pr(J_i = f \mid (Z_i, X_i) = (z^\ell, x))} \\ &\quad - \frac{E(Y_{fi} \mid (Z_i, X_i) = (z^\ell, x), J_i = f) \Pr(J_i = f \mid (Z_i, X_i) = (z^\ell, x))}{\Pr(J_i = f \mid (Z_i, X_i) = (z^h, x)) - \Pr(J_i = f \mid (Z_i, X_i) = (z^\ell, x))} \end{aligned} \tag{5.14}$$

for $\varphi\left(z^\ell, x\right) < \varphi(z^h, x)$. Now notice that for any ν,

$$\begin{aligned} &\lim_{\varphi(z^\ell, x) \uparrow \nu, \varphi(z^h, x) \downarrow \nu} E(Y_{fi} \mid \varphi(z^\ell, x) \leq \nu_i < \varphi(z^h, x), X_i = x) \\ &\quad = E(Y_{fi} \mid \nu_i = \nu, X_i = x). \end{aligned}$$

Thus if (x, ν) is in the support of $(X_i, \varphi(Z_i, X_i))$, then $E(Y_{fi} \mid \nu_i = \nu, X_i = x)$ is identified. Since the model is symmetric, under similar conditions $E(Y_{hi} \mid \nu_i = \nu,$

$X_i = x$) is identified as well. Finally since

$$\begin{aligned}\Delta^{\text{MTE}}(x, \nu) &= E(\pi_i \mid X_i = x, \nu_i = \nu) \\ &= E(Y_{fi} \mid \nu_i = \nu, X_i = x) - E(Y_{hi} \mid \nu_i = \nu, X_i = x), \end{aligned} \tag{5.15}$$

the marginal treatment effect is identified.

The marginal treatment effect is interesting in its own right. It is the value of the treatment for any individual with $X_i = x$ and $\nu_i = \nu$. In addition, it is also useful because the different types of treatment effects can be defined in terms of the marginal treatment effect. For example

$$\text{ATE} = \int \int_0^1 \Delta^{\text{MTE}}(x, \nu) \text{d}\nu \text{d}G(x).$$

One can see from this expression that without full support this will not be identified because $\Delta^{\text{MTE}}(x, \nu)$ will not be identified everywhere.

Heckman and Vytlacil (2005) also show that the instrumental variables estimator defined in Eq. (5.5) (conditional on x) is

$$\int_0^1 \Delta^{\text{MTE}}(x, \nu) h_{IV}(x, \nu) \text{d}\nu$$

where they give an explicit functional form for h_{IV}. It is complicated enough that we do not repeat it here but it can be found in Heckman and Vytlacil (2005).

This framework is also useful for seeing what is not identified. In particular if $\varphi(Z_i, x)$ does not have full support so that it is bounded above or below, the average treatment effect will not be identified. However, many other interesting treatment effects can be identified. For example, the Local Average Treatment Effect in a model with no regressors (x) is

$$\text{LATE} = \frac{\int_{\varphi(0)}^{\varphi(1)} \Delta^{\text{MTE}}(\nu) \text{d}\nu}{\varphi(1) - \varphi(0)}. \tag{5.16}$$

More generally, in this series of papers, Heckman and Vytlacil show that the marginal treatment effect can also be used to organize many ideas in the literature. One interesting case is policy effects. They define the policy relevant treatment effect as the treatment resulting from a particular policy. They show that if the relationship between the policy and the observable covariates is known, the policy relevant treatment effect can be identified from the marginal treatment effects.

5.4. Applications of the marginal treatment effects approach

Heckman and Vytlacil (1999, 2001, 2005) suggest procedures to estimate the marginal treatment effect. They suggest what they call "local instrumental variables." Using our notation for the generalized Roy model in which $J_i = f$ when $\varphi(X_i, Z_i) - \nu_i > 0$, where ν_i is uniformly distributed, they show that

$$\Delta^{\text{MTE}}(x, \nu) = \frac{\partial E(Y_i \mid X_i = x, \varphi(X_i, Z_i) = \nu)}{\partial \nu}.$$

To see why this is the same definition of MTE as in Eq. (5.15)), note that

$$\begin{aligned}
&\frac{\partial E(Y_i \mid X_i = x, \varphi(X_i, Z_i) = \nu)}{\partial \nu} \\
&= \frac{\partial \left[E(Y_{fi} \mid X_i = x, \nu_i \le \nu) \Pr(\nu_i \le \nu) + E(Y_{hi} \mid X_i = x, \nu_i > \nu) \Pr(\nu_i > \nu) \right]}{\partial \nu} \\
&= \frac{\partial \left[\int_0^\nu E(Y_{fi} \mid \nu_i = \omega, X_i = x) \mathrm{d}\omega + \int_\nu^1 E(Y_{hi} \mid \nu_i = \omega, X_i = x) \mathrm{d}\omega \right]}{\partial \nu} \\
&= E(Y_{fi} \mid \nu_i = \nu, X_i = x) - E(Y_{hi} \mid \nu_i = \nu, X_i = x) \\
&= \Delta^{\text{MTE}}(x, \nu).
\end{aligned}$$

Thus one can estimate the marginal treatment effect in three steps. First estimate φ, second estimate $E(Y_i \mid X_i = x, \varphi(X_i, Z_i) = \nu)$ using some type of nonparametric regression approach, and third take the derivative.

Because as a normalization ν_i is uniformly distributed

$$\begin{aligned}
\varphi(x, z) &= \Pr(\nu_i \le \varphi(X_i, Z_i) \mid X_i = x, Z_i = z) \\
&= \Pr(J_i = f \mid X_i = x, Z_i = z) \\
&= E(D_{fi} \mid X_i = x, Z_i = z).
\end{aligned}$$

Thus we can estimate $\varphi(x, z)$ from a nonparametric regression of D_{fi} on (X_i, Z_i).

A very simple way to do this is to use a linear probability model of D_{fi} regressed on a polynomial of Z_i. By letting the terms in the polynomial get large with the sample size, this can be considered a nonparametric estimator. For the second stage we regress the outcome Y_i on a polynomial of our estimate of $\varphi(Z_i)$. To see how this works consider the case in which both polynomials are quadratics. We would use the following two stage least squares procedure:

$$D_{fi} = \gamma_0 + \gamma_1 Z_i + \gamma_2 Z_i^2 + \gamma_x X_i + e_i, \tag{5.17}$$

$$Y_i = \beta_0 + \beta_1 \widehat{D_{fi}} + \beta_2 \widehat{D_{fi}}^2 + \beta_x X_i + u_i, \tag{5.18}$$

where $\widehat{D_{fi}} = \widehat{\gamma_0} + \widehat{\gamma_1} Z_i + \widehat{\gamma_2} Z_i^2 + \widehat{\gamma_x} X_i$ is the predicted value from the first stage. The β_2 coefficient may not be 0 because as we change $\widehat{D_{fi}}$ the instrument affects different groups of people. The MTE is the effect of changing $\widehat{D_{fi}}$ on Y_i. For the case above the MTE is:

$$\frac{\partial Y_i}{\partial \widehat{D_{fi}}} = \beta_1 + 2\beta_2 \widehat{D_{fi}}. \tag{5.19}$$

Although the polynomial procedure above is transparent, the most common technique used to estimate the MTE is local linear regression.

French and Song (2010) estimate the labor supply response to Disability Insurance (DI) receipt for DI applicants. Individuals are deemed eligible for DI benefits if they are "unable to engage in substantial gainful activity"—i.e., if they are unable to work. Beneficiaries receive, on average $12,000 per year, plus Medicare health insurance. Thus, there are strong incentives to apply for benefits. They continue to receive these benefits only if they earn less than a certain amount per year ($10,800 in 2007). For this reason, the DI system likely has strong labor supply disincentives. A healthy DI recipient is unlikely to work if that causes the loss of DI and health insurance benefits.

The DI system attempts to allow benefits only to those who are truly disabled. Many DI applicants have their case heard by a judge who determines those who are truly disabled. Some applicants appear more disabled than others. The most disabled applicants are unable to work, and thus will not work whether or not they get the benefit. For less serious cases, the applicant will work, but only if she is denied benefits. The question, then, is what is the optimal threshold level for the amount of observed disability before the individual is allowed benefits? Given the definition of disability, this threshold should depend on the probability that an individual does not work, even when denied the benefit. Furthermore, optimal taxation arguments suggest that benefits should be given to groups whose labor supply is insensitive to benefit allowance. Thus the effect of DI allowance on labor supply is of great interest to policy makers.

OLS is likely to be inconsistent because those who are allowed benefits are likely to be less healthy than those who are denied. Those allowed benefits would have had low earnings even if they did not receive benefits. French and Song propose an IV estimator using the process of assignment of cases to judges. Cases are assigned to judges on a rotational basis within each hearing office, which means that for all practical purposes, judges are randomly assigned to cases conditional on the hearing office and the day. Some judges are much more lenient than others. For example, the least lenient 5% of all judges allow benefits to less than 45% of the cases they hear, whereas the most lenient 5% of all judges allow benefits to 80% of all the cases they hear. Although some of those

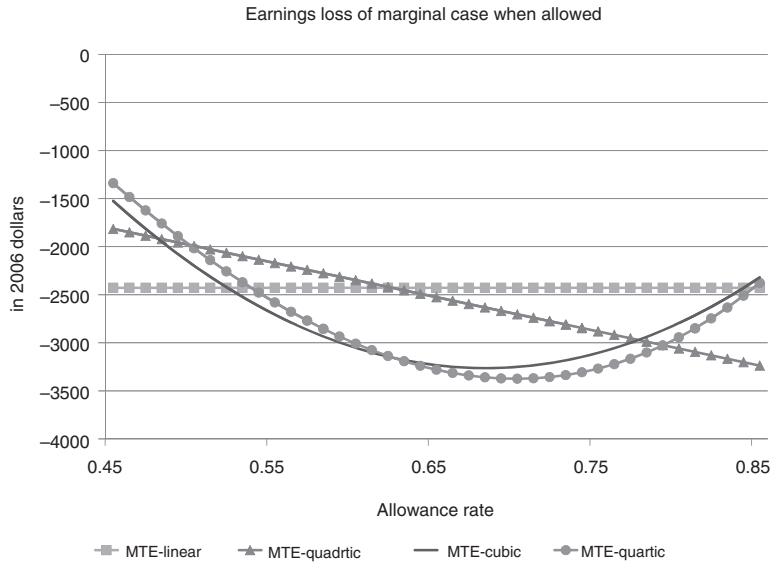


Figure 1 ***Marginal treatment effect.***

who are denied benefits appeal and get benefits later, most do not. If assignment of cases to judges is random then the instrument of judge assignment is a plausibly exogenous instrument. Furthermore, and as long as judges vary in terms of leniency and not ability to detect individuals who are disabled,[15] the instrument can identify a MTE.

French and Song use a two stage procedure. In the first stage they estimate the probability that an individual is allowed benefits, conditional on the average judge specific allowance rate. They estimate a version of Eq. (5.17) where D_{fi} is an indicator equal to 1 if case i was allowed benefits and Z_i is the average allowance rate of the judge who heard case i. In the second stage they estimate earnings conditional on whether the individual was allowed benefits (as predicted by the judge specific allowance rate). They estimate a version of Eq. (5.18) where Y_i is annual earnings 5 years after assignment to a judge. Figure 1 shows the estimated MTE (using the formula in Eq. (5.19)) using several different specifications of polynomial in the first and second stage equations. Assuming that the treatment effect is constant (i.e., $\beta_2 = 0$), they find that annual earnings 5 years after assignment to a judge are \$1500 for those allowed benefits and \$3900 for those denied benefits, so the estimated treatment effect is \$2400. This is the MTE-linear case in Fig. 1. However, this masks considerable heterogeneity in the treatment effects. They find that when allowance rates rise, the labor supply response of the marginal case also rises. When allowing for the quadratic term β_2 to be non-zero, they find that less lenient

[15] If judges vary in terms of ability to detect disability, then a case that is allowed by a low allowance judge might be denied by a high allowance judge. This would violate the monotonicity assumption shown in Eq. (5.7).

judges (who allow 45% of all cases) have a MTE of a $1800 decline in earnings. More lenient judges (who allow 80% of all cases) have a MTE of $3200 decline in earnings. Figure 1 also shows results when allowing for cubic and quartic terms in the polynomials in the first and second stage equations. This result is consistent with the notion that as allowance rates rise, more healthy individuals are allowed benefits. These healthier individuals are more likely to work when not receiving DI benefits, and thus their labor supply response to DI receipt is greater.

One problem with an instrument such as this is that the instrument lacks full support. Even the most lenient judge does not allow everyone benefits. Even the strictest judge does not deny everyone. However, the current policy debate is whether the thresholds should be changed by only a modest amount. For this reason, the MTE on the support of the data is the effect of interest, whereas the ATE is not.

Doyle (2007) estimates the Marginal Treatment Effect of foster care on future earnings and other outcomes. Foster care likely increases earnings of some children but decreases it for others. For the most serious child abuse cases, foster care will likely help the child. For less serious cases, the child is probably best left at home. The question, then, is at what point should the child abuse investigator remove the child from the household? What is the optimal threshold level for the amount of observed abuse before which the child is removed from the household and placed into foster care?

Only children from the most disadvantaged backgrounds are placed in foster care. They would have had low earnings even if they were not placed in foster care. Thus, OLS estimates are likely inconsistent. To overcome this problem, Doyle uses IV. Case investigators are assigned to cases on a rotational basis, conditional on time and the location of the case. Case investigators are assigned to possible child abuse cases after a complaint of possible child abuse is made (by the child's teacher, for example). Investigators have a great deal of latitude about whether the child should be sent into foster care. Furthermore, some investigators are much more lenient than others. For example, one standard deviation in the case manager removal differential (the difference between his average removal rate and the removal rate of other investigators who handle cases at the same time and place) is 10%. Whether the child is removed from the home is a good predictor of whether the child is sent to foster care. So long as assignment of cases to investigators is random and investigators only vary in terms of leniency (and not ability to detect child abuse) then the instrument of investigator assignment is a useful and plausibly exogenous instrument.

Doyle uses a two stage procedure where in the first stage he estimates the probability that a child is placed in foster care as a function of the investigator removal rate. In the second stage he estimates adult earnings as a function of whether the child was placed in foster care (as predicted by the instrument). He finds that children placed into foster care earn less than those not placed into foster care over most of the range of the data. Two stage least squares estimates reveal that foster care reduces adult quarterly earnings

by about \$1000, which is very close to average earnings. Interestingly, he finds that when child foster care placement rates rise, earnings of the marginal case fall. For example, earnings of the marginal child handled by a lenient investigator (who places only 20% of the children in foster care) are unaffected by placement. For less lenient investigators, who place 25% of the cases in foster care, earnings of the marginal case decline by over \$1500.

Carneiro and Lee (2009) estimate the counterfactual marginal distributions of wages for college and high school graduates, and examine who enters college. They find that those with the highest returns are the most likely to attend college. Thus, increases in college cause changes in the distribution of ability among college and high school graduates. For fixed skill prices, they find that a 14% increase in college participation (analogous to the increase observed in the 1980s) reduces the college premium by 12%. Likewise, Carneiro et al. (2010) find that while the conventional IV estimate of the return to schooling (using distance to a college and local labor market conditions as the instruments) is 0.095, the estimated marginal return to a policy that expands each individual's probability of attending college by the same proportion is only 0.015.

5.5. Selection on observables

Perhaps the simplest and most common assumption is that assignment of the treatment is random conditional on observable covariates (sometimes referred to as unconfoundedness). The easiest way to think about this is that the selection error term is independent of the other error terms:

Assumption 5.2.

$$J_i = f \quad \text{when } \varphi(X_i) > \nu_i$$

where ν_i is independent of $(\varepsilon_{fi}, \varepsilon_{hi})$.

We continue to assume that $Y_{fi} = g_f(X_{fi}, X_{0i}) + \varepsilon_{fi}$ and $Y_{hi} = g_h(X_{hi}, X_{0i}) + \varepsilon_{hi}$. Note that we have explicitly dropped Z_i from the model as we consider cases in which we do not have exclusion restrictions. The implication of this assumption is that unobservable factors that determine one's income as a fisherman do not affect the choice to become a fisherman. That is while it allows for selection on observables in a very general way, it does not allow for selection on unobservables.

Interestingly, this is still not enough for us to identify the Average Treatment Effect. If there are values of observable covariates X_i for which $\Pr(J_i = f \mid X_i = x) = 1$ or $\Pr(J_i = f \mid X_i = x) = 0$ the model is not identified. If $\Pr(J_i = f \mid X_i = x) = 1$ then it is straightforward to identify $E(Y_{fi} \mid X_i = x)$, but $E(Y_{hi} \mid X_i = x)$ is not identified. Thus we need the additional assumption

Assumption 5.3. For almost all x in the support of X_i,

$$0 < \Pr(J_i = f \mid X_i = x) < 1.$$

Theorem 5.2. *Under Assumptions* 5.2 and 5.3 *the Average Treatment Effect is identified.*

(Proof in Appendix.)

Estimation in this case is relatively straightforward. One can use matching[16] or regression analysis to estimate the average treatment effect.

5.6. Set identification of treatment effects

In our original discussion of identification we defined $\Psi(\Theta(P))$ as "the set of values of ψ that are consistent with the data distribution P." We said that ψ was identified if this set was a singleton. However, there is another concept of identification we have not discussed until this point; this is set identification. Sometimes we may be interested in a parameter that is not point identified, but this does not mean we cannot say anything about it. In this subsection we consider the case of set identification (i.e. trying to characterize the set $\Psi(\Theta(P))$) focusing on the case in which ψ is the Average Treatment Effect. Suppose that we have some prior knowledge (possibly an exclusion restriction that gives us a LATE). What can we learn about the ATE without making any functional form assumptions? In a series of papers Manski (1989, 1990, 1995, 1997) and Manski and Pepper (2000, 2009) develop procedures to derive set estimators of the Average Treatment Effect and other parameters given weak assumptions. By "set identification" we mean the set of possible Average Treatment Effects given the assumptions placed on the data. Throughout this section we will continue to assume that the structure of the Generalized Roy model holds and we derive results under these assumptions. In many cases the papers we mentioned do not impose this structure and get more general results.

Following Manski (1990) or Manski (1995), notice that

$$E\left(Y_{fi}\right) = E(Y_{fi} \mid J_i = f)\Pr(J_i = f) + E(Y_{fi} \mid J_i = h)\Pr(J_i = h) \quad (5.20)$$

$$E\left(Y_{hi}\right) = E(Y_{hi} \mid J_i = h)\Pr(J_i = h) + E(Y_{hi} \mid J_i = f)\Pr(J_i = f). \quad (5.21)$$

We observe all of the objects in Eqs (5.20) and (5.21) except $E(Y_{fi} \mid J_i = h)$ and $E(Y_{hi} \mid J_i = f)$. The data are completely uninformative about these two objects. However, suppose we have some prior knowledge about the support of Y_{fi} and Y_{hi}. In particular, suppose that the support of Y_{fi} and Y_{hi} are bounded above by y^u and from below by y^ℓ. Thus, by assumption $y^u \geq E(Y_{fi} \mid J_i = h) \geq y^\ell$ and $y^u \geq E(Y_{hi} \mid J_i = f) \geq y^\ell$.

[16] Our focus is on identification rather than estimation. Thus we avoid a discussion of matching estimators. See Heckman et al. (1999), Imbens and Wooldridge (2009), or DiNardo and Lee (2011) for discussion.

Using these assumptions and Eqs (5.20) and (5.21) we can establish that

$$\begin{aligned} &E(Y_{fi} \mid J_i = f)\Pr(J_i = f) + y^{\ell}\Pr(J_i = h) \\ &\quad \leq E\left(Y_{fi}\right) \leq E(Y_{fi} \mid J_i = f)\Pr(J_i = f) + y^{u}\Pr(J_i = h) \qquad (5.22) \\ &E(Y_{hi} \mid J_i = h)\Pr(J_i = h) + y^{\ell}\Pr(J_i = f) \\ &\quad \leq E\left(Y_{hi}\right) \leq E(Y_{hi} \mid J_i = h)\Pr(J_i = h) + y^{u}\Pr(J_i = f). \qquad (5.23) \end{aligned}$$

Using these bounds and the definition of the ATE

$$\text{ATE} = E\left(Y_{fi}\right) - E\left(Y_{hi}\right) \qquad (5.24)$$

yields

$$\begin{aligned} &(E(Y_{fi} \mid J_i = f)\Pr(J_i = f) + y^{\ell}\Pr(J_i = h)) \\ &\quad - (E(Y_{hi} \mid J_i = h)\Pr(J_i = h) + y^{u}\Pr(J_i = f)) \\ &\leq \text{ATE} \\ &\leq (E(Y_{fi} \mid J_i = f)\Pr(J_i = f) + y^{u}\Pr(J_i = h)) \\ &\quad - (E(Y_{hi} \mid J_i = h)\Pr(J_i = h) + y^{\ell}\Pr(J_i = f)). \end{aligned}$$

In practice the bounds above can yield wide ranges and are often not particularly informative. A number of other assumptions can be used to decrease the size of the identified set.

Manski (1990, 1995) shows that one method of tightening the bounds is with an instrumental variable. We can write the expressions (5.20) and (5.21) conditional on $Z_i = z$ for any $z \in \text{supp}(Z_i)$ as for each $j \in \{f, h\}$,

$$\begin{aligned} E\left(Y_{ji}|Z_i = z\right) = {} & E(Y_{ji} \mid J_i = f, Z_i = z)\Pr(J_i = f \mid Z_i = z) \\ & + E(Y_{ji} \mid J_i = h, Z_i = z)\Pr(J_i = h \mid Z_i = z). \qquad (5.25) \end{aligned}$$

Since Z_i is, by assumption, mean independent of Y_{fi} and Y_{hi} (it only affects the probability of choosing one occupation versus the other), then $E\left(Y_{fi}|Z_i = z\right) = E\left(Y_{fi}\right)$ and $E\left(Y_{hi}|Z_i = z\right) = E(Y_{hi})$. Assume there is a binary instrumental variable, Z_i, which equals either 0 or 1. We can then follow exactly the same argument as in Eqs (5.22) and (5.23), but conditioning on Z_i and using Eq. (5.25) yields

$$\begin{aligned} &E(Y_{fi} \mid J_i = f, Z_i = 1)\Pr(J_i = f \mid Z_i = 1) + y^{\ell}\Pr(J_i = h \mid Z_i = 1) \\ &\quad \leq E\left(Y_{fi}\right) \\ &\quad \leq E(Y_{fi} \mid J_i = f, Z_i = 1)\Pr(J_i = f \mid Z_i = 1) + y^{u}\Pr(J_i = h \mid Z_i = 1) \qquad (5.26) \end{aligned}$$

$$
\begin{aligned}
&E(Y_{hi} \mid J_i = h, Z_i = 0)\Pr(J_i = h \mid Z_i = 0) + y^{\ell}\Pr(J_i = f \mid Z_i = 0) \\
&\quad \leq E\,(Y_{hi}) \\
&\quad \leq E(Y_{hi} \mid J_i = h, Z_i = 0)\Pr(J_i = h \mid Z_i = 0) + y^{u}\Pr(J_i = f \mid Z_i = 0). \quad (5.27)
\end{aligned}
$$

Thus we can bound $\text{ATE} = E(Y_{fi}) - E(Y_{hi})$ from below by subtracting (5.27) from (5.26):

$$
\begin{aligned}
&E(Y_{fi} \mid J_i = f, Z_i = 1)\Pr(J_i = f \mid Z_i = 1) + y^{\ell}\Pr(J_i = h \mid Z_i = 1) \\
&\qquad - E(Y_{hi} \mid J_i = h, Z_i = 0)\Pr(J_i = h \mid Z_i = 0) + y^{u}\Pr(J_i = f \mid Z_i = 0) \\
&\quad \leq \text{ATE} \\
&\quad \leq E(Y_{fi} \mid J_i = f, Z_i = 1)\Pr(J_i = f \mid Z_i = 1) \\
&\qquad + y^{u}\Pr(J_i = h \mid Z_i = 1) - E(Y_{hi} \mid J_i = h, Z = 0) \\
&\qquad \times \Pr(J_i = h \mid Z_i = 0) + y^{\ell}\Pr(J_i = f \mid Z_i = 0). \quad (5.28)
\end{aligned}
$$

Our choice of a binary value of Z_i can be trivially relaxed. In the cases in which Z_i takes on many values one could choose any two values in the support of Z_i to get upper and lower bounds. If our goal is to minimize the size of the set we would choose the values z^{ℓ} and z^{h} to minimize the difference between the upper and lower bounds in (5.28):

$$
(y^{u} - y^{\ell})[\Pr(J_i = h \mid Z_i = z^{h}) + \Pr(J_i = f \mid Z_i = z^{\ell})].
$$

The importance of support conditions once again becomes apparent from this expression. If we could find values z^{ℓ} and z^{h}such that

$$
\begin{aligned}
\Pr(J_i = h \mid Z_i = z^{h}) &= 0 \\
\Pr(J_i = f \mid Z_i = z^{\ell}) &= 0
\end{aligned}
$$

then this expression is zero and we obtain point identification of the ATE. When $\Pr(J_i = h \mid Z_i = z)$ or $\Pr(J_i = f \mid Z_i = z)$ are bounded from below we are only able to obtain set estimates. A nice aspect of this is that it represents a nice middle point between identifying LATE versus claiming the ATE is not identified. If the identification at infinity effect is not exactly true, but approximately true so that one can find values of z^{ℓ} and z^{h} so that $\Pr(J_i = h \mid Z_i = z^{h})$ and $\Pr(J_i = f \mid Z_i = z^{\ell})$ are small, then the bounds will be tight. If one cannot find such values, the bounds will be far apart.

In many cases these bounds may be wide. Wide bounds can be viewed in two ways. One interpretation is that the bounding procedure is not particularly helpful in learning about the true ATE. However, a different interpretation is that it shows that the

data, without additional assumptions, is not particularly helpful for learning about the ATE. Below we discuss additional assumptions for tightening the bounds on the ATE, such as Monotone treatment response, Monotone treatment selection, and Monotone instruments. In order to keep matters simple, below we assume that there is no exclusion restriction. However, if a exclusion restriction is known, this allows us to tighten the bounds.

Next we consider the assumption of Monotone Treatment Response introduced in Manski (1997), which we write as

Assumption 5.4. Monotone Treatment Response

$$Y_{fi} \geq Y_{hi}$$

with probability one.

In the fishing/hunting example this is not a particularly natural assumption, but for many applications in labor economics it is. Suppose we are interested in knowing the returns to a college degree, and Y_{fi} is income for individual i if a college graduate whereas Y_{hi} is income if a high school graduate. It is reasonable to believe that the causal effect of school or training cannot be negative. That is, one could reasonably assume that receiving more education can't causally lower your wage. Thus, Monotone Treatment Response seems like a reasonable assumption in this case. This can lower the bounds above quite a bit because now we know that

$$E(Y_{fi} \mid J_i = h) \geq E(Y_{hi} \mid J_i = h) \tag{5.29}$$
$$E(Y_{hi} \mid J_i = f) \leq E(Y_{fi} \mid J_i = f). \tag{5.30}$$

From this Manski (1997) shows that

$$0 \leq \text{ATE}.$$

Another interesting assumption that can also help tighten the bounds is the Monotone Treatment Selection assumption introduced in Manski and Pepper (2000). In our framework this can be written as

Assumption 5.5. Monotone Treatment Selection: for $j = f$ or h,

$$E(Y_{ji} \mid J_i = f) \geq E(Y_{ji} \mid J_i = h).$$

Again this might not be completely natural for the fishing/hunting example, but may be plausible in many other cases. For example it seems like a reasonable assumption in schooling if we believe that there is positive sorting into schooling. Put differently,

suppose the average college graduate is a more able person than the average high school graduate and would earn higher income, even if she did not have the college degree. If this is true, then the average difference in earnings between college and high school graduates overstates the true causal effect of college on earnings. This also helps to further tighten the bounds as this implies that

$$\text{ATE} \leq E(Y_{fi} \mid J_i = f) - E(Y_{hi} \mid J_i = h).$$

Note that by combining the MTR and MTS assumption, one can get the tighter bounds:

$$0 \leq \text{ATE} \leq E(Y_{fi} \mid J_i = f) - E(Y_{hi} \mid J = h).$$

Manski and Pepper (2000) also develop the idea of a monotone instrumental variable. An instrumental variable is defined as one for which for any two values of the instrument z_a and z_b,

$$E(Y_{ji} \mid Z_i = z_a) = E(Y_{ji} \mid Z_i = z_b).$$

In words, the assumption is that the instrument does not directly affect the outcome variable Y_{ji}. It only affects one's choices. Using somewhat different notation, but their exact wording, they define a monotone instrumental variable in the following way

Assumption 5.6. Let $\mathcal{Z}$ be an ordered set. Covariate Z_i is a monotone instrumental variable in the sense of mean-monotonicity if, for $j \in \{f, h\}$,each value of x, and all $(z_b, z_a) \in (\mathcal{Z} \times \mathcal{Z})$ such that $z_b \geq z_a$,

$$E(Y_{ji} \mid X_i = x, Z_i = z_b) \geq E(Y_{ji} \mid X_i = x, Z_i = z_a).$$

This is a straight generalization of the instrumental variable assumption, but imposes much weaker requirements for an instrument. It does not require that the instrument be uncorrelated with the outcome, but simply that the outcome monotonically increase with the instrument. An example is that parental income has often been used as an instrument for education. Richer parents are better able to afford a college degree for their child. However, it seems likely that the children of rich parents would have had high earnings, even in the absence of a college degree.

They show that this implies that

$$\sum_{z \in \mathcal{Z}} \Pr(Z_i = z) \left\{ \sup_{z_a \leq z} \left[E\left(Y_i \mid Z_i = z_a, J_i = f\right) \Pr\left(J_i = f \mid Z_i = z_a\right) \right.\right.$$
$$\left.\left. + \; y^{\ell} \Pr\left(J_i = h \mid Z_i = z_a\right) \right] \right\}$$

$$
\begin{aligned}
&- \sum_{z \in \mathcal{Z}} \Pr(Z_i = z) \left\{ \inf_{z_b \geq z} \left[E\left(Y_i \mid Z_i = z_b, J_i = h\right) \Pr\left(J_i = h \mid Z_i = z_b\right) \right.\right. \\
&\left.\left. + \; y^u \Pr\left(J_i = f \mid Z_i = z_b\right) \right] \right\} \\
&\leq \text{ATE} \\
&\leq \sum_{z \in \mathcal{Z}} \Pr(Z_i = z) \left\{ \inf_{z_b \geq z} \left[E\left(Y_i \mid Z_i = z_b, J_i = f\right) \Pr\left(J_i = f \mid Z_i = z_b\right) \right.\right. \\
&\left.\left. + \; y^u \Pr\left(J_i = h \mid Z_i = z_b\right) \right] \right\} \\
&- \sum_{z \in \mathcal{Z}} \Pr(Z_i = z) \left\{ \sup_{z_a \leq z} \left[E\left(Y_i \mid Z_i = z_a, J_i = h\right) \Pr\left(J_i = h \mid Z_i = z_a\right) \right.\right. \\
&\left.\left. + \; y^\ell \Pr\left(J_i = f \mid Z_i = z_a\right) \right] \right\}.
\end{aligned}
$$

One can obtain tighter bounds by combining the Monotone Instrumental Variable assumption with the Monotone Treatment Response assumption but we do not explicitly present this result.

Blundell et al. (2007) estimate changes in the distribution of wages in the United Kingdom using bounds to allow for the impact of non-random selection into work. They first document the growth in wage inequality among workers over the 1980s and 1990s. However, they point out that rates of non-participation in the labor force have grown in the UK over the same time period. Nevertheless, they show that selection effects alone cannot explain the rise in inequality observed among workers: the worst case bounds establish that inequality has increased. However, worst case bounds are not sufficiently informative to understand such questions as whether most of the rise in wage inequality is due to increases in wage inequality within education groups versus across education groups. Next, they add an additional assumptions to tighten the bounds. First, they assume the probability of work is higher for those with higher wages, which is essentially the Monotone Treatment Selection assumption shown in Assumption 5.5. Second, they make the Monotone Instrumental Variables assumption shown in Assumption 5.6. They assume that higher values of out of work benefit income are positively associated with wages. They show that both of these assumptions tighten the bounds considerably. They find that when these additional restrictions are made, then they can show that both within group and between group inequality has increased.

5.7. Using selection on observables to infer selection on unobservables

Altonji et al. (2005a) suggest another approach which is to use the amount of selection on observable covariates as a guide to the potential amount of selection on unobservables.

To motivate this approach, consider an experiment in which treatment status is randomly assigned. The key to random assignment is that it imposes that treatment status be independent of the unobservables in the treatment model. Since they are unobservable, one can never explicitly test whether the treatment was truly random. However, if randomization was carried out correctly, treatment should also be uncorrelated with observable covariates. This is testable, and applying this test is standard in experimental approaches.

Researchers use this same argument in non-experimental cases as well. If a researcher wants to argue that his instrument or treatment is approximately randomly assigned, then it should be uncorrelated with observable covariates as well. Even if this is strictly not required for consistent estimates of instrumental variables, readers may be skeptical of the assumption that the instrument is uncorrelated with the unobservables if it is correlated with the observables. Researchers often test for this type of relationship as well.[17] The problem with this approach is that simply testing the null of uncorrelatedness is not that useful. Just because you reject the null does not mean it isn't approximately true. We would not want to throw out an instrument with a tiny bias just because we have a data set large enough to detect a small correlation between it and an observable. Along the same lines, just because you fail to reject the null does not mean it is true. If one has a small data set with little power one could fail to reject the null even though the instrument is poor. To address these issues, Altonji et al. (2005a) design a framework that allows them to describe how large the treatment effect would be if "selection on the unobservables is the same as selection on the observables."

Their key variables are discrete, so they consider a latent variable model in which a dummy variable for graduation from high school can be written as

$$G_i = \begin{cases} 1 & Y_i^* \geq 0 \\ 0 & Y_i^* < 0 \end{cases}$$

where Y_i^* can be written as

$$\begin{aligned} Y_i^* &= \beta_0 + \alpha D_{fi} + \sum_{j=1}^{K} W_{ij}\beta_j \\ &= \beta_0 + \alpha D_{fi} + \sum_{j=1}^{K} S_j W_{ij}\beta_j + \sum_{j=1}^{K}(1 - S_j) W_{ij}\beta_j \\ &= \beta_0 + \alpha D_{fi} + X_i'\beta + \nu_i . \end{aligned}$$

W_{ij} represent all covariates, both those that are observable to the econometrician and those that are unobservable, the variable S_j is a dummy variable representing whether the

[17] Altonji et al. (2005a) discuss a number of studies that do so.

covariate is observable to the empirical researcher, $X_i'\beta = \sum_{j=1}^{K} S_j W_{ij}\beta_j$ represents the observable part of the index, and $v_i = \sum_{j=1}^{K}(1 - S_j)W_{ij}\beta_j$ denotes the unobservable part.

Within this framework, one can see that different assumptions about what dictates which observables are chosen (S_j) can be used to identify the model. Their specific goal is to quantify what it means for "selection on the observables to be the same as selection on the unobservables." They argue that the most natural way to formalize this idea is to assume that S_j is randomly assigned so that the unobservables and observables are drawn from the same underlying distribution.

The next question is what this assumption implies on the data that can be useful for identification. They consider the projection:

$$\text{proj}(Z_i \mid X_i'\beta, v_i) = \phi_0 + \phi X_i'\beta + \phi_\varepsilon v_i$$

where Z_i can be any random variable. They show that if S_j is randomly assigned,

$$\phi \approx \phi_\varepsilon.$$

This restriction is typically sufficient to insure identification of α.[18]

Altonji et al. (2005a,b) argue that for their example this is an extreme assumption and the truth is somewhere in between this assumption and the assumption that Z_i is uncorrelated with the unobservables which would correspond to $\phi_\varepsilon = 0$. They assume that when $\phi > 0$,

$$0 \leq \phi_\varepsilon \leq \phi.$$

There are at least three arguments for why selection on unobservables would be expected to be less severe than selection on observables (as it is measured here). First, some of the variation in the unobservable is likely just measurement in the dependent variable. Second, data collectors likely collect the variables that are likely to be correlated with many things. Third, there is often a time lapse between the time the baseline data is collected (the observables) and when the outcome is realized. If unanticipated events occur in between these two time periods, that would lead to the result.

Notice that if $\phi = 0$ then assuming $\phi_\varepsilon = \phi$ is the same as assuming $\phi_\varepsilon = 0$. However, if ϕ were very large the two estimates would be very different, which would shed doubt on the assumption of random assignment. Since ϕ essentially picks up the relationship between the instrument and the observable covariates, the bounds would be wide when

[18] In some cases it is not point identification, but either 2 or 3 different points.

there is a lot of selection on observables and will be tight when there is little selection on observables.

Altonji, Elder, and Taber consider the case of whether the decision to attend Catholic high school affects outcomes such as test scores and high school graduation rates. Those who attend Catholic schools have higher graduation rates than those who do not attend Catholic schools. However, those who attend Catholic may be very different from those who do not. They find that (on the basis of observables) while this is true in the population, it is not true when one conditions on the individuals who attend Catholic school in eighth grade. To formalize this, they use their approach and estimate the model under the two different assumptions. In their application the projection variable, Z_i, is the latent variable determining whether an individual attends Catholic school. First they estimate a simple probit of high school graduation on Catholic high school attendance as well as many other covariates. This corresponds to the $\phi_\varepsilon = 0$ case. They find a marginal effect of 0.08, meaning that Catholic school raises high school graduation by eight percentage points. Next they estimate a bivariate probit of Catholic high school attendance and high school graduation subject to the constraint that $\phi_\varepsilon = \phi$. In this case they find a Catholic high school effect of 0.05. The closeness of these two estimates strongly suggests that the Catholic high school effect is not simply a product omitted variable bias. The tightness of the two estimates arose both because ϕ was small and because they use a wide array of powerful explanatory variables.

6. DURATION MODELS AND SEARCH MODELS

In this section we relate the previous discussion to the competing risks model and the search model. We show that the competing risk model can be written in a way that is almost identical to the Roy model. We also show how the basic ideas of exclusion restrictions can be used to identify a version of a search model.

6.1. Competing risks model

With duration data a researcher observes the elapsed time until some event occurs. The prototypical example in labor economics is the duration of unemployment and we focus on that example. We explain why identification of this model is almost identical to identification of the Roy model. Let T_i denote the length of an unemployment spell. There are (at least) four different ways to characterize the distribution of T_i. The first is the cumulative distribution function $F(t) \equiv \Pr(t > T_i)$, which in the context of unemployment durations is the probability the individual found a job. The second is the density function f. The third is the survivor function defined as

$$S(t) \equiv \Pr(T_i > t) = 1 - F(t).$$

The fourth is the hazard function, which is the job finding rate at time t, given that the individual was unemployed at time t:

$$\begin{aligned} h(t) &\equiv \lim_{\delta \to 0} \frac{\Pr(T_i \leq t + \delta \mid T_i \geq t)}{\delta} \\ &= \frac{f(t)}{S(t)}. \end{aligned}$$

The link between the hazard rate and survivor function is:

$$\begin{aligned} h(t) = \frac{f(t)}{S(t)} &= \frac{\mathrm{d}F(t)/\mathrm{d}t}{S(t)} \\ &= \frac{-\mathrm{d}S(t)/\mathrm{d}t}{S(t)} \\ &= \frac{-\mathrm{d}\log S(t)}{\mathrm{d}t}. \end{aligned} \tag{6.1}$$

There is a large literature on identification of duration models. Heckman and Taber (1994), Van den Berg (2001), and Abbring (2010) provide excellent surveys of this literature.[19] Rather than survey the full literature here we relate it to our previous discussion. Given that T_i must be positive, it is natural to model T_i using the basic framework we have been using all along:

$$\log(T_i) = g(X_i) + \varepsilon_i.$$

Clearly if we could observe the distribution of $\log(T_i)$ conditional on X_i, identification of g and the distribution of ε_i would be straightforward.

However, often we cannot observe the full duration of T_i because the spell (or our observation of it) is truncated before the worker is re-employed. For example, the worker may die, be lost from the data, or the survey may end. In the classic medical example we might want to estimate the duration until a patient has a heart attack, but if she dies from cancer we never observe this event. Hence the name "competing risk model." To put this in the context of our Roy model example, suppose an unemployed worker would take the first offer they received and they can get an offer as a fisherman or a hunter. Define the model as

$$\log(T_{fi}) = g_f(X_i) + \varepsilon_{fi} \tag{6.2}$$

$$\log(T_{hi}) = g_h(X_i) + \varepsilon_{hi} \tag{6.3}$$

[19] Key papers include Elbers and Ridder (1982), Heckman and Singer (1984a,b), Ridder (1990), Honoré (1993), and Abbring and Ridder (2009).

where T_{fi} and T_{hi} are the amount of time it would take until the worker received an offer as a fisherman or as a hunter, X_i denotes observable variables that are independent of the unobservables $(\varepsilon_{fi}, \varepsilon_{hi})$.[20] The econometrician can observe whether the worker becomes a fisherman or a hunter and the length of the unemployment spell. However, notice that as Heckman and Honoré (1990) point out, this is just another version of the Roy model. Rather than observe the maximum of Y_{fi} and Y_{hi}, the econometrician observes the minimum of $\log(T_{fi})$ and $\log(T_{hi})$.

The specification (6.2) and (6.3) above is not the way that many researchers choose to model duration data. Often they model the hazard function directly as it is sometimes easier to interpret. Moreover, if the observable covariates change over time, the hazard model is a more reasonable way to model the durations. The most common specification is the mixed proportional hazard model

$$h(t \mid X_i = x) = \xi(t)\phi(x)\omega_i \tag{6.4}$$

where $\xi(t)$ is referred to as the baseline hazard, ω_i is an unobservable variable which is independent of the observables, and X_i denotes observable characteristics. Most studies find that the hazard rate for finding a job tends to decline with the unemployment duration. The model above allows for two possible interpretations of this empirical regularity. First, it could be that as unemployment durations lengthen, skills depreciate, making it harder to find a job. This is captured by $\xi(t)$. Second, it could be that some people are just less able to find a job than others in ways not captured by observables. This is captured in ω_i. Van den Berg (2001) provides a thorough discussion of this model.

Heckman and Honoré (1989) show how to map the hazard specification into a framework that is similar to what we use in our analysis of the Roy model. The transformation is simplest is when $\xi(t) = 1$. In that case one can write the survivor function as

$$\Pr(T_i > t \mid X_i = x, \omega_i = \omega) = \mathrm{e}^{-t\phi(x)\omega}. \tag{6.5}$$

It is straightforward to derive Eq. (6.4) using the survivor function (6.5) and Eq. (6.1). Define $g(\cdot) = -\log(\phi(\cdot))$ and F_ω to be the distribution of ω_i. In order to obtain the cumulative density function of unemployment durations we must integrate over the distribution of unemployed individuals:

$$\Pr(T_i \leq t \mid X_i = x) = \int 1 - \mathrm{e}^{-t\phi(x)\omega_i} \mathrm{d}F_\omega$$

[20] We do not need to make use of exclusion restrictions here so we do not distinguish between observables that may enter differently.

$$= \int 1 - \exp(-\exp(\log(t) - g(x) + \log(\omega_i)))\mathrm{d}F_\omega$$
$$\equiv F_{\widetilde{\omega}}(\log(t) - g(x)) \tag{6.6}$$

where $F_{\widetilde{\omega}}$ is defined implicitly by this relationship. Note that $F_{\widetilde{\omega}}$ is a legitimate CDF, as it is strictly increasing from 0 to 1.[21] Thus one can think of the data generating process as

$$\log(T_i) = g(X_i) + \widetilde{\omega}_i$$

where $\widetilde{\omega}_i$ is distributed according to $F_{\widetilde{\omega}}$ and is independent of X_i.

In the more general case in which $\xi(t)$ is not constant, it is well known that one can write the survivor function as

$$\mathrm{e}^{-\Xi(t)\phi(X_i)\omega_i} \tag{6.7}$$

where Ξ is the integrated hazard

$$\Xi(t) \equiv \int_0^t \xi(t)\mathrm{d}t.$$

Equation (6.7) differs from Eq. (6.5) by the term $\Xi(t)$ instead of t. Thus replacing t with $\Xi(t)$ in Eq. (6.6) yields

$$\log(\Xi(T_i)) = g(X_i) + \widetilde{\omega}_i.$$

Heckman and Honoré (1989) use a more general framework to think about the competing risks model in which the probability of not getting a fishing job by time t_f and not getting a hunting job by time t_h, $S(t_f, t_h \mid X_i = x)$, can be written as

$$S(t_f, t_h \mid X_i = x) = K(\exp\{-\Xi_f(t_f)\phi_f(x)\}, \exp\{-\Xi_h(t_h)\phi_h(x)\})$$

where $\phi_j(x) = \exp(-g_j(x))$ for $j = f, h$. This is a generalization of a model in which

$$\log(\Xi_f(T_{fi})) = g_f(X_i) + \widetilde{\omega}_{fi}$$
$$\log(\Xi_h(T_{hi})) = g_h(X_i) + \widetilde{\omega}_{hi}$$

because

$$S(t_f, t_h \mid X_i = x) = \Pr[\log(\Xi_f(T_{fi})) > \log(\Xi_f(t_f)), \log(\Xi_h(T_{hi}))$$
$$> \log(\Xi_h(t_h)) \mid X_i = x]$$

[21] It is the distribution of a convolution between $\log(\omega_i)$ and an extreme value.

$$
\begin{aligned}
&= \Pr[g_f(x) + \widetilde{\omega}_{fi} > \log(\Xi_f(t_f)), g_h(x) + \widetilde{\omega}_{hi} > \log(\Xi_h(t_h))] \\
&= \Pr[-\widetilde{\omega}_{fi} < -\log(\Xi_f(t_f)) + g_f(x), -\widetilde{\omega}_{hi} \\
&\quad < -\log(\Xi_h(t_h)) + g_h(x)] \\
&= F_{-\widetilde{\omega}_{fi}-\widetilde{\omega}_{hi}}(-\log(\Xi_f(t_f)) + g_f(x), -\log(\Xi_h(t_h)) + g_h(x)) \\
&\equiv K(\exp\{-\Xi_f(t_f)\phi_f(x)\}, \exp\{-\Xi_h(t_h)\phi_h(x)\}) \qquad (6.8)
\end{aligned}
$$

where $F_{-\widetilde{\omega}_{fi}-\widetilde{\omega}_{hi}}$ is the joint CDF of $(-\widetilde{\omega}^*_{fi}, -\widetilde{\omega}^*_{hi})$, and K is defined implicitly as $K(a, b) = F_{-\widetilde{\omega}_{fi}-\widetilde{\omega}_{hi}}(-\log(-\log(a)), -\log(-\log(b)))$.

Heckman and Honoré (1989), Theorem 1 contains the following result. We reproduce their result, only altering the notation.

Theorem 6.1. *Assume that (T_{fi}, T_{hi}) has the joint survivor function as given in (6.8). Then $\Xi_f, \Xi_h, \phi_f, \phi_h$, and K are identified from the identified minimum of (T_{fi}, T_{hi}) under the following assumptions*

1. *K is continuously differentiable with partial derivatives K_1 and K_2 for $i = 1, 2$, the limit as $n \to \infty$ of $K_i(\eta_{1n}, \eta_{2n})$ is finite for all sequences of η_{1n}, η_{2n} for which $\eta_{1n} \to 1$ and $\eta_{2n} \to 1$ for $n \to \infty$. We also assume that K is strictly increasing in each of its arguments in all of $[0, 1] \times [0, 1]$.*
2. *$\Xi_f(1) = 1, \Xi_h(1) = 1, \phi_f(x_0) = 1$ and $\phi_h(x_0) = 1$ for some fixed point x_0 in the support X.*
3. *The support of $\{\phi_f(x), \phi_h(x)\}$ is $(0, \infty) \times (0, \infty)$.*
4. *Ξ_f and Ξ_h are nonnegative, differentiable, strictly increasing functions, except that we allow them to be ∞ for finite t.*

(Proof in Heckman and Honoré (1989).)

Since the model is almost identical to the Roy model, the intuition for identification is very similar so we don't review it here. We do mention a few things about these assumptions. First note that assumption (2) in Theorem 6.1 is just a normalization as one cannot separate the scales of ϕ_f, Ξ_f, and ν_f. The more notable difference between this and the theorem we presented in the Roy model section above is the lack of exclusion restrictions. What is crucial in being able to do this is the assumptions about K in assumption (1). In their proof they show that for any x in the support of X_i,

$$
\lim_{t \to 0} \frac{\frac{\partial \Pr(T_{fi} < t, T_{hi} > T_{fi} | X_i = x)}{\partial t}}{\frac{\partial \Pr(T_{fi} < t, T_{hi} > T_{fi} | X_i = x_0)}{\partial t}} = \phi_f(x).
$$

One could in principle use this form of identification for the Roy model, but it is somewhat less natural in the Roy framework, as taking the limit as $t \to 0$ corresponds to

taking limits as the log of wages become arbitrarily large. It also makes heavy use of the independence assumption, which is not necessary for identification of g_f when one has exclusion restrictions. Finally, the basic approach will not expand to the "labor supply" model in which we only observe wages in one sector and to the generalized Roy model in the same way that exclusion restrictions do.

Abbring and van den Berg (2003) extends Heckman and Honoré's (1989) results on the mixed proportional hazards competing risk models in a few ways, including generalizing the assumptions for identification somewhat and considering identification in the case in which researchers observe multiple spells.

6.2. Search models

Eckstein and van den Berg (2007) present a nice survey of Empirical Search models. We avoid a general discussion, but rather combine the proportional hazard model with a search model. In a well known result Flinn and Heckman (1982) show that the search model is not fully identified. They use the Lippman and McCall (1976) search model in which workers search for jobs until their wage exceeds their reservation wage. In this model, one essentially assumes that the worker stays at the job forever. All workers are assumed to be ex-ante identical and face the same distribution of offered wages, which we denote by F. The reservation wage w^r is the point at which the individual is indifferent between taking the job and continued search. It is defined implicitly by the formula

$$c + w^r = \frac{\lambda}{r} \int_{w^r}^{\infty} (x - w^r) \mathrm{d}F(x)$$

where c is search cost, r is the interest rate, and λ is the hazard rate of finding a job.

Flinn and Heckman (1982) assume that one observes the time until finding a job (T_i) and the wage a worker receives conditional on finding the job. The only source of heterogeneity in the model comes from the timing of the job offers and the draw from the wage offer distribution. Clearly one can identify the distribution of accepted wage offers which is the distribution of observed wages. The reservation wage is the lowest acceptable wage, so one can identify w^r as the minimum observed wage. Then they can identify

$$\frac{f(x)}{1 - F(w^r)} \quad \text{for } x \geq w^r.$$

They can also identify the hazard rates of job finding which is

$$\lambda(1 - F(w^r)).$$

However, this is all that can be identified. In particular, one cannot separate λ from $(1 - F(w^r))$. Furthermore, the distribution of wage offers below the reservation wage is not identified. This is quite intuitive. Since nobody works at a salary below the reservation wage, we do not have any information from the data on what that distribution might look like.[22] Furthermore, identification of the model above relies on the strong assumption that people are identical. All dispersion in observed wages comes from identical people with identical skills being offered different wages. It also implies a constant hazard rate of finding jobs λ, which is at odds with the data.

By using exclusion restrictions and using some of the ideas from the Roy model with the arguments from the mixed proportional hazard model, most of the components of the model can be identified. In particular let the arrival rate of job offers be

$$\lambda_i = \phi(X_{\lambda i}, X_{0i})\omega_i \tag{6.9}$$

where now $X_{\lambda i}$ is an exclusion restriction that influences the arrival rate, but not any other aspect of the model. We assume that search cost is defined as

$$\log(C_i) = g_h(X_{hi}, X_{0i}) + \varepsilon_{hi}. \tag{6.10}$$

Finally we assume the wage offer that individual i would receive at time t is

$$\log(W_{fit}) = g_f(X_{fi}, X_{0i}) + \varepsilon_{fit}. \tag{6.11}$$

The complicated aspect of this model is that workers may reject the first offer they receive, and then receive a second different offer. Thus we need the time subscript on ε_{fit} to denote that this draw can be different. The second issue is that one would expect the distribution of offered ε_{fit} to not be identical across workers. We assume that the distribution of ε_{fit} is individual specific coming from distribution $F_{i\varepsilon_f}$. That is each time a worker gets a new offer it is a draw from the distribution of $F_{i\varepsilon_f}$. As above X_i is observable and independent of $(\nu_i, \varepsilon_{fit}, \varepsilon_{hi})$.

Using the Lippman and McCall (1976) model, define W_i^* as the solution to the equation

$$C_i + W_i^* = \frac{\lambda_i}{r} \int_{\log(W_i^*) - g_f(X_{fi}, X_{0i})}^{\infty} (\mathrm{e}^{g_f(X_{fi}, X_{0i}) + \varepsilon_{fit}} - W_i^*) \mathrm{d}F_{i\varepsilon_f}(\varepsilon_{fit}). \tag{6.12}$$

[22] Of course this raises an interesting question. What does it mean for a firm to make an offer that it knows no worker would ever take? In most wage posting models, a firm would never post a wage that no worker would take (see e.g. Burdett and Mortensen, 1998). However, if there is a job match component, one can also write down a model in which one could define the counterfactual wage at which a worker would be paid at a job in which he would never take (whether that offer is actually "extended" or not is largely a semantic issue).

The reservation wage is defined as

$$W_i^r \equiv \max\{W_i^*, 0\}. \tag{6.13}$$

If search costs are sufficiently high, W_i^* could be negative. But because the distribution of wages is bounded below at 0, the reservation wage would be 0.

The added assumptions to identify the model are completely analogous to those we used for the Roy model earlier

Assumption 6.1. $(\varepsilon_{fit}, \varepsilon_{hi}, \nu_i)$ is continuously distributed with support $\mathbb{R}^3$, and is independent of X_i.

Assumption 6.2. $\text{supp}(\phi(X_{\lambda i}, X_{0i}), g_f(X_{fi}, x_0), g_h(X_{hi}, x_0)) = \mathbb{R}^+ \times \mathbb{R}^2$ for all $x_0 \in \text{supp}(X_{0i})$.

Assumption 6.3. The marginal distributions of ε_{fit},ε_{hi}, and ν_i have expected values equal to zero. Moreover, the expected value of $e^{\varepsilon_{fit}}$ is finite.

Assumption 6.4. $X_i = (X_{fi}, X_{hi}, X_{\lambda i}, X_{0i})$ can be written as $(X_{fi}^c, X_{fi}^d, X_{hi}^c, X_{hi}^d, X_{\lambda i}^c, X_{\lambda i}^d, X_{0i}^c, X_{0i}^d)$ where the elements of $X^c = (X_{fi}^c, X_{hi}^c, X_{\lambda i}^c, X_{0i}^c)$ are continuously distributed (no point has positive mass), and $X^d = (X_{fi}^d, X_{fi}^d, X_{\lambda i}^d, X_{0i}^d)$ is distributed discretely (all support points have positive mass).

Assumption 6.5. For any $(x_f^d, x_h^d, x_\lambda^d, x_0^d) \in \text{supp}(X_{fi}^d, X_{hi}^d, X_{\lambda i}^d, X_{0i}^d)$, $g_f(x_f^c, x_f^d, x_0^c, x_0^d)$, $g_h(x_h^c, x_h^d, x_0^c, x_0^d)$, and $\phi(x_\lambda^c, x_\lambda^d, x_0^c, x_0^d)$ are almost surely continuous across $(x^c) \in \text{supp}(X_i^c \mid X_i^d = x^d)$.

Theorem 6.2. *Under Assumptions 6.1–6.5 and that ϕ and the distribution of ω_i satisfy the assumptions in Heckman and Honoré (1989), given that we observe T_i and w_{fiT_i} from the model determined by Eqs (6.9)–(6.13), we can identify ϕ and g_f on their support, and g_h up to location on a set $\mathcal{X}^*$ that has measure 1.*

(Proof in Appendix.)

Unlike some of the other models, we have not completely identified the error structure (or the location of g_h). This is probably not surprising given the complexity of $F_{i\varepsilon_f}$ and the relatively modest data conditions.[23]

[23] Some aspects of the distribution of wages can be identified. For example identification of the marginal distribution of ω_i is straightforward. Describing the distribution of $F_{i\varepsilon_f}$ is difficult because it is a distribution of distributions. Given the cost in setting up notation to discuss this, we do not try to characterize this distribution. A typical assumption would be that we could write $\varepsilon_{fit} = \epsilon_{fi} + \zeta_{fit}$, where ϵ_{fi} is an individual specific term that does not vary across wages and ζ_{fit} is iid.

We conclude this section after making three comments. First, it is not clear that one cares about the location of g_h. That is, for many interesting policy counterfactuals, identification of the aspects above should be sufficient. Second, with more structure, more features of the model should be identified.[24] Third, if a researcher observes multiple spells on the same worker, this can add much identifying information. The identification problem arises because if we see one worker making more than another we do not know if it is because the first worker is more productive or if they just happened to get a fortunate draw from offer distribution. With panel data, if we see that the first worker consistently earns more money across many employers, this would suggest that the difference has more to do with ability than with draws from the offer distribution.

We have barely scratched the surface of identification of search models. Many papers being estimated today are based on equilibrium models such as Mortensen and Pissarides (1994), Burdett and Mortensen (1998), or Postel-Vinay and Robin (2002). We think there is much work to be done on identification in these models.[25]

7. FORWARD LOOKING DYNAMIC MODELS

In this section we discuss an extension of the generalized Roy model into a dynamic framework with uncertainty and forward looking behavior. We show that the basic identification ideas presented above can be generalized to dynamic models. The identification results for the simple models on which we focus can be extended to more complicated environments. We begin with a two period model in which there are three choices made over two periods. We then discuss some general issues with identifying the components of the Bellman Equation. Finally we present a dynamic Generalized Roy model that one can use for dynamic treatment effect evaluation. Once again, we do not provide a full review of the literature, but focus on expanding the generalized Roy model into a forward looking dynamic model. Abbring (2010) includes a more complete discussion.[26]

7.1. Two period discrete choice dynamic model

We begin with the framework of Taber (2000) who considers a simple version of a dynamic model. To think of this model as an extension of the basic Roy model we go from two occupational choices to three. While we could modify the fishing/hunting example to a dynamic context, it is easiest to think about this in terms of an education

[24] Proving identification in nonlinear models such as this one is often quite difficult. This might not be problematic in practice as researchers can search for multiple solutions in the data. If there are multiple solutions, all can be reported. If only one solution exists, this should give a consistent estimate of the truth.

[25] Canals-Cerda (2010) provides a recent example which adds measurement error in wages to the Flinn and Heckman (1982) framework. Barlevy (2008) shows how to non-parametrically identify the wage offer distribution in the presence of measurement error in wages and unobserved heterogeneity in skills.

[26] Recent papers that cover aspects of identification not discussed here include Kashara and Shimotsu (2009) and Hu and Shum (2009).

model as Taber does. In particular, a student first decides whether to graduate from high school or not. After graduating from high school, she decides whether to attend college or enter the labor market directly. Extending beyond 3 choices is straightforward, but as in Taber we stick to the 3 choice model for expositional purposes. We focus on identification of the choice model and ignore data on earnings until Section 7.3.

First consider the case in which there was no uncertainty or dynamics. We specify the model using the three value functions

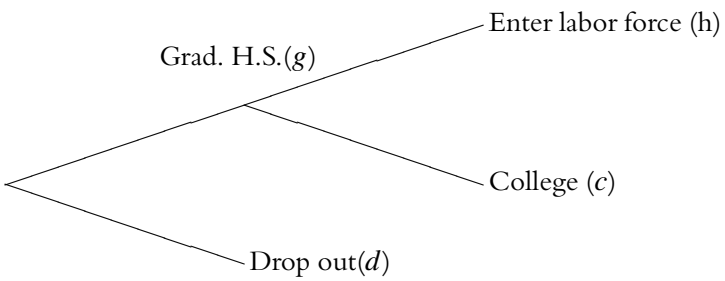
$$\begin{aligned} V_{ci} &= g_c(X_{ci}, X_{0i}) + \varepsilon_{ci} \\ V_{di} &= g_d(X_{di}, X_{0i}) + \varepsilon_{di} \\ V_{hi} &= 0 \end{aligned}$$

where V_{ci} is the value function for a college student, V_{hi} the value function for an individual with exactly a high school degree, and V_{di} the value function for high school dropout. Individuals choose the option with the highest value function. That is

$$J_i = \text{argmax}\{V_{di}, V_{hi}, V_{ci}\}.$$

If there were no uncertainty in this model it would be a simple polychotomous choice model. Matzkin (1993) considers identification a general class of polychotomous choice modes under a number of different assumptions. One result is that since choices are only identified up to monotonic transformations, $V_{hi} = 0$ is a location normalization that we impose at this point. Adding dynamics and uncertainty does not change this result.

Our goal now is to add dynamics and uncertainty to the model. The timing can be seen in the following figure



In the first period the agent chooses whether to graduate from high school. If she graduates in the first period, she then chooses whether to go to college in the second.

The key aspect of the model is that information will be revealed between the first and second period. The agent's preferences are summarized by lifetime reward function V_{ji} at each terminal state $j \in \{c, h, d\}$. Taber defines V_{di} so that it is known at the time the high school graduation choice is made. Then in period two, V_{ci} and V_{hi} are known when the choice between c and h is made. That is, in period one the agent does not know X_{ci} or ε_{ci}. The first period information is assumed to be contained in $(X_{0i}, X_{1i}, \varepsilon_{1i})$, where X_{1i} is observable in period one and will be informative about X_{ci} while ε_{1i} is unobservable and informative about ε_{ci}. We assume that decisions are made in order to maximize expected lifetime reward. Thus the reward function at node g in the first period takes the value

$$V_g(x_1, x_d, x_0, \epsilon_1) \equiv E[\max\{V_{ci}, V_{hi}\} \mid (X_{1i}, X_{di}, X_{0i}) = (x_1, x_d, x_0), \varepsilon_{1i} = \epsilon_1].$$

The agent chooses node d if $V_{di} > V_g(X_{1i}, X_{di}, X_{0i}, \varepsilon_{1i})$ and chooses node g otherwise. If she chooses g in the first period she chooses node c in the second if $V_{ci} > V_{hi}$ and node h otherwise.

We let $G(X_{ci} \mid (X_{1i}, X_{di}, X_{0i}) = (x_1, x_d, x_0))$ denote the distribution of X_{ci} conditional on $(X_{1i}, X_{di}, X_{0i}) = (x_1, x_d, x_0)$. We can summarize the information structure as follows

Known to the Agent at time one	**Learned by the Agent at time two**	**Observed by the Econometrician**
$\varepsilon_{1i}, \varepsilon_{di}$	ε_{ci}	X_{0i}, X_{1i}, X_{di}
X_{0i}, X_{1i}, X_{di}	X_{ci}	X_{ci}
$G(X_{ci} \mid (X_{1i}, X_{di}, X_{0i}) = (x_1, x_d, x_0))$		J_i

We first consider identification of g_c and g_d up to monotonic transformations. We follow Taber (2000) closely except that we use our notation and use stronger assumptions than he does to avoid adding more notation.[27]

Assumption 7.1. For any $(x_c, x_0) \in \text{supp}\{X_{ci}, X_{0i}\}$,

$$\begin{aligned} \text{supp}\{\varepsilon_{di}\} &= \mathbb{R} = \text{supp}\{g_d(X_{di}, x_0) \mid (X_{ci}, X_{0i}) = (x_c, x_0)\} \\ \text{supp}\{\varepsilon_{ci}\} &= \mathbb{R}. \end{aligned}$$

This assumption is analogous to what we have been assuming all along. In order to estimate the full model, we need full support of g_d conditional on (X_{ci}, X_{0i}).

[27] Taber (2000) allows for the possibility that the support of the error term could be bounded, which allows for weaker support condition on the observables.

Assumption 7.2. For any $(x_d, x_0) \in \text{supp}\{X_{di}, X_{0i}\}$, $y \in \mathbb{R}$, and $a \in (0,1)$, there exists a set $\mathcal{X}_1(x_f, x_0, y, a)$ with positive measure such that for $x_1 \in \mathcal{X}_1(x_f, x_0, y, a)$,

(a) $\Pr(g_c(X_{ci}, x_0) < y \mid (X_{1i}, X_{di}, X_{0i}) = (x_1, x_d, x_0)) > a$.
(b) The distribution of $g_c(X_{ci}, x_0)$ conditional on $(X_{1i}, X_{di}, X_{0i}) = (x_1, x_d, x_0)$ is stochastically dominated by the unconditional distribution of $g_c(X_{ci}, x_0)$.

This is a stochastic analogue of a support condition. In the case in which X_{ci} were known at time one so that $X_{1i} = X_{ci}$, this would be implied to be a standard support condition. However, it is general enough to allow for the distribution of X_{ci} to not be known at time one, but we still need a time one variable X_{1i} that is useful in forecasting X_{ci}. For example X_{ci} could be a variable like family income while the child is in college while X_{1i} is a variable like family income while the child is in high school. This assumption states that we can condition on the value of this variable so that the conditional probability that the agent chooses option c in the second period can become arbitrarily small. In the family income example this means we could condition on families whose income while the child is in high school are sufficiently low that college seems like a very unlikely outcome for the child.

Assumption 7.3. $(\varepsilon_{1i}, \varepsilon_{di}, \varepsilon_{ci})$ is independent of $(X_{1i}, X_{di}, X_{ci}, X_{0i})$, for any $\epsilon_1 \in \text{supp}(\varepsilon_{1i})$,

$$E(|\varepsilon_{ci}| \mid \varepsilon_{1i} = \epsilon_1) < \infty$$

and for any $(x_1, x_d, x_0) \in \text{supp}(X_{1i}, X_{di}, X_{0i})$,

$$E\left(|g_c(X_{ci}, x_0)| \mid (X_{1i}, X_{di}, X_{0i}) = (x_1, x_d, x_0)\right) < \infty.$$

Assumption 7.3 is the separable independent assumption that we have been making throughout this chapter. We also need to assume that the stochastic components have finite expectations so that V_g is finite.

Theorem 7.1. *Under Assumptions 7.1–7.3, from data on $(X_{1i}, X_{di}, X_{ci}, X_{0i}, J_i)$, g_d and g_c are identified up to monotonic transformation.*

(Proof in Taber (2000).)

The basic strategy used in this proof is a stochastic extension of "identification at infinity." This should not be surprising as this looks very much like the type of selection problem we have discussed throughout this chapter: we can not observe the choice between c and h unless individuals have already rejected d.

We identify g_c in almost exactly the same way as we identified g_f as presented for the Roy Model. With an exclusion restriction we can condition on g_d arbitrarily low so that the probability of selecting node d is close to zero. This leaves us with a simple

binary choice model in which the agents choose between h and c. The type of exclusion restriction used here is a variable that enters g_d, but does not influence g_c directly. One can see this in the following expression

$$\begin{aligned}
&\lim_{g_d(x_d,x_0)\to-\infty} \Pr(J_i = c \mid X_i = x) \\
&= \lim_{g_d(x_d,x_0)\to-\infty} \Pr[g_d(x_d, x_0) + \varepsilon_{di} \leq V_g(x_1, x_d, x_0, \varepsilon_{1i}), g_c(x_c, x_0) + \varepsilon_{ci} > 0] \\
&= \Pr[g_c(x_c, x_0) + \varepsilon_{ci} > 0].
\end{aligned}$$

Using standard identification strategies for the binary choice model described in the first step of identification of the Roy model, g_c is identified.

Identification of g_d is somewhat trickier, but one can use essentially the same idea. In a static model one could use an identification at infinity argument by eliminating c as an option and could compare the binary choice of d versus h. In this stochastic case this is can not be done because the value of X_{ci} is not known at time 1. Thus we need a somewhat different type of exclusion restriction, a variable known at time one that does not enter g_d directly, but does have predictive power for the distribution of g_c above and beyond X_{di}. To see how this works, suppose we have a variable X_{1i} that satisfies these conditions and that as x_1 gets small the conditional distribution of g_c shifts to the left. In this case

$$\lim_{x_1\to-\infty} E\left[\max\left(g_c(X_{ci}, x_0) + \varepsilon_{ci}, 0\right) \mid (X_{1i}, X_{di}, X_{0i}) = (x_1, x_d, x_0), \varepsilon_{1i} = \epsilon_1\right] = 0,$$

so that

$$\begin{aligned}
&\lim_{x_1\to-\infty} \Pr(J_i = d \mid X_i = x) \\
&\quad= \lim_{x_1\to-\infty} \Pr\left[g_d(x_d, x_0) + \varepsilon_{di} > E\left[\max\left(V_{ci}, 0\right) \mid (X_{1i}, X_{di}, X_{0i})\right.\right. \\
&\qquad\left.\left.= (x_1, x_d, x_0), \varepsilon_{1i} = \epsilon_1\right]\right] \\
&\quad= \Pr[g_d(x_d, x_0) + \varepsilon_{di} > 0].
\end{aligned}$$

From this piece we can identify g_d up to a monotonic transformation. This type of variable will satisfy Assumption 7.2. Note that the type of exclusion restriction we need here is something that is known at time 1, is useful in forecasting X_{ci}, but does not affect V_{di}.

Taber (2000) goes on to consider identification of the distribution of the error terms. The most general version of the full model above can not be identified without further assumptions, so he instead studies a few interesting cases. Identification of the error terms requires a different kind of exclusion restriction. His key assumption requires variation in $g_c(x_c)$ holding x_1 fixed. Thus we need some uncertainty from the point of view of the agents. The full model is not identified if agent's have perfect information about future

values of X_{ci}. A natural way to satisfy this exclusion restriction is with time varying observables. The details can be found in Taber (2000).

7.2. Identification of the components of the Bellman equation

While the model above is dynamic, we have not used Bellman's equation. A natural way to parameterize the model would be to define period specific utility functions $u_h(X_{hi}, X_{0i}, \varepsilon_{hi})$, $u_c(X_{ci}, X_{0i}, \varepsilon_{ci})$, and $u_g(X_{1i}, X_{0i}, \varepsilon_{1i})$ in each of the three nodes above other than the dropout node. If we think of the model as a two period model we can define $u_d(t, X_{di}, X_{0i}, \varepsilon_{di})$ to be the period specific utility of individual i if she drops out at time t. Conditional on graduating, she enters college if

$$u_c(X_{ci}, X_{0i}, \varepsilon_{ci}) > u_h(X_{hi}, X_{0i}, \varepsilon_{hi}).$$

The Bellman equation for the high school graduate is

$$\begin{aligned} V_g(x_1, x_d, x_0, \epsilon_1) \equiv u_g(x_1, x_0, \epsilon_1) + \beta E[\max\{&u_c(X_{ci}, X_{0i}, \varepsilon_{ci}), \\ &u_h(X_{hi}, X_{0i}, \varepsilon_{hi})\} \mid (X_{1i}, X_{di}, X_{0i}) = (x_1, x_d, x_0), \varepsilon_{1i} = \epsilon_1]. \end{aligned}$$

Mapping back to the notation in the subsection above, the rest of the value functions are defined as

$$\begin{aligned} V_{di} &= u_d(1, X_{di}, X_{0i}, \varepsilon_{di}) + \beta u_d(2, X_{di}, X_{0i}, \varepsilon_{di}) \\ V_{hi} &= u_g(X_{1i}, X_{0i}, \varepsilon_{1i}) + \beta u_h(X_{hi}, X_{0i}, \varepsilon_{hi}) \\ V_{ci} &= u_g(X_{1i}, X_{0i}, \varepsilon_{1i}) + \beta u_c(X_{ci}, X_{0i}, \varepsilon_{ci}). \end{aligned}$$

An obvious question arises as to whether one can separately identify the components of the value functions β, u_h, u_c, and u_d. Unfortunately, in general one can not do this. Consider a full certainty version of the model. In this case the decision of which occupation to enter would depend on V_{di}, V_{hi}, and V_{ci} only. One can choose any $\beta > 0$ and any u_g, but then always find a value of u_c and u_h to leave V_{ci} and V_{hi} unchanged. For a simple model such as the one Taber (2000) presents, parameterizing the model in terms of the terminal value functions (i.e. V_{di}, V_{hi}, and V_{ci}) avoids this problem as one does not need to decompose them into their components.

However, Taber's parameterization is clearly not feasible for an infinitely lived model. Furthermore, it is not convenient in an finite time model with many periods and state variables. It does not take advantage of the dimension reducing advantages of the Bellman formulation: the functions would depend on the whole history of state variables rather than just the current set.

Next we consider Rust's (1994) model. Note that we use his notation exactly even though it is inconsistent with our previous notation. Let S_i represent the current state

and D_i represent the discrete choice. In general S_i will contain elements that are both observed and unobserved by the econometrician. He writes the Bellman equation as

$$v(s,d) = u(s,d) + \beta \int \max_{D_i' \in D(S_i')} [v(S_i', D_i')] p(dS_i' \mid S_i = s, D_i = d)$$

where v is the value function, u is the period specific utility function, β is the discount rate, $D(s)$ is the choice set in state of the world s, and p is the transitional probability distribution of the state variables. Rust (1994) shows that one can not separately identify the model above from an alternative with the same β and p, but with

$$\bar{u}(s,d) = u(s,d) + f(s) - \beta \int f(S_i') p(ds' \mid S_i = s, D_i = d).$$

Intuitively this is close to the discussion above in the simple model in which you can change the timing at which the innovation to utility takes place, without changing the value function.

Magnac and Thesmar (2002) discuss this issue in much greater detail. They not only show that the model is not identified, but document the extent of underidentification. They additionally assume that one can write

$$u(S_i, d) = u_d(X_i) + \varepsilon_{di}$$

where X_i is the observable part of the state space and the unobservable ε_{di} is mean independent of x and independent across periods (conditional on x and d). That is S_i represents the state space, so if one knows S_i, they also know X_i and ε_{di}. They show that given knowledge of β and the joint distribution of the ε_{di}, one can identify

$$U_d(x) \equiv u_d(x) + \beta \int \max_{D_i' \in D(D_i')} [v(S_i', D_i')] p(dS_i' \mid X_i = x, D_i = d) - u_k(x)$$
$$+ \beta \int \max_{D_i' \in D(S_i')} [v(S_i', D_i')] p(dS_i' \mid X_i = x, D_i = k)$$

where k is one of the elements of $D(s)$. They further explore the model with additional identifying information and correlated random effects.

How problematic is it that the model is not fully identified? The answer to this question depends on the purpose of the model. That is, even if the model is not fully identified, one may still be able to identify policy counterfactuals of interest. Ichimura and Taber (2002) provide one example of a case in which the policy counterfactual can be identified. They start with the model of Keane and Wolpin (2001) and show how

one can estimate a semiparametric reduced form version of this model and use it to evaluate the effect of a tuition subsidy on college enrollment. They key is having enough structure on the model to map variation in the data to the counterfactual tuition subsidy.

Aguirregabiria (2010) presents a different and somewhat more general example of policy evaluation in a finite time dynamic discrete choice model. We do not get into the details as it is different from the types of labor models we study here, but he shows that, despite the fact that his full model is not identified, the welfare effect function resulting from the policy change can be identified. Thus one can do welfare analysis even though the full model is not identified.

7.3. Dynamic generalized Roy model

Heckman and Navarro (2007) provide another example showing that one can identify interesting counterfactuals even when the full model is not identified. Their study complements the discussion in this chapter as it extends the work on identification in dynamic discrete choice models into the treatment effects literature discussed in Section 5 above. They consider a finite time optimal stopping problem. Using the notation used above in Section 7.2, D_i is either zero or one, and once it is one it remains one forever. Their main example is a schooling model in which students decide at which time to leave school (assuming that after leaving they cannot come back). The model is essentially a dynamic generalized Roy model. Let T_{ia} and L_{ia} respectively denote the level of schooling and a dummy for whether individual i is out of school at age a. Using a somewhat modified version of their notation we can write time a earnings as

$$Y_{i,a,t,\ell} = \mu(a, t, \ell, X_i) + \varepsilon_{i,a,t,\ell}$$

where t and ℓ represent potential outcomes of $T_{i,a}$ and $L_{i,a}$. Heckman and Navarro (2007) also assume that the cost of schooling can be written as

$$C_{i,t} = \Phi(t, X_i, Z_i) + \omega_{i,t}.$$

In order to keep our notation complete and consistent across sections we will assume that random variable $\Theta_{i,a}$ summarizes all information (both observables and unobservables) that individual i has at age a. This means that if we know $\Theta_{i,a}$ we also know $(X_i, Z_i, T_{i,a}, L_{i,a}, \varepsilon_{i,a,t,l}, \omega_{i,t})$, so when we condition on $\Theta_{i,a} = \theta$, we are conditioning on $(X_i, Z_i, T_{i,a}, L_{i,a}, \varepsilon_{i,a,t,l}, \omega_{i,t}) = (x, z, t, \ell, \epsilon_{a,t,\ell}, \omega_t)$. We will make use of this notation below.

Once a student leaves school they make no further decisions, so if a student leaves school at age a with t years of schooling, lifetime utility discounted to the time one

leaves school is written as

$$R(a,t,\theta) = E\left(\sum_{j=0}^{\bar{T}}\left(\frac{1}{1+r}\right)^{j} Y_{i,a+j,t,1} \mid \Theta_{i,a} = \theta\right).$$

The only decision that agents make is whether they will drop out of school or not. For a student at age a with t years of schooling the value function when they make this decision is written as

$$V(a,t,\theta) = \max\left\{R(a,t,\theta), \mu(a,t,0,x) + \epsilon_{a,t,0} - \Phi(t,x,z) - \omega_t + \left(\frac{1}{1+r}\right) E\left[V(a+1,t+1,\Theta_{i,a+1}) \mid \Theta_{i,a} = \theta\right]\right\}.$$

This is basically a dynamic version of the generalized Roy model. Identification follows by essentially combining the arguments used by Taber (2000) for the dynamic aspects of the model with the arguments for identification of the generalized Roy model. Heckman and Navarro (2007) use higher level assumptions to avoid the use of exclusion restrictions.[28] They also use a factor structure on the distribution of the error term to reduce dimension. We refer readers interested in these generalizations and in the details of their proof to their paper. Here we attempt to give an intuitive feel for identification of this model and show how it is related to identification of the generalized Roy model presented in Section 3.3.

Identification of reduced form choice model

In this case they do not derive an explicit reduced form, but note that

$$\Pr(T_{i,a} = t \mid X_i = x, Z_i = z)$$

can be identified directly from the data.

Identification of the earnings equation μ

With exclusion restrictions this can be done in exactly the same way as in the static model. Assuming that $\varepsilon_{i,a,t,\ell}$ has a zero mean,

$$\lim_{\Pr(T_{i,a}=t|(X_i,Z_i)=(x,z))\to 1} E\left[Y_{i,a+j,t,1} \mid (X_i,Z_i) = (x,z)\right] = \mu(a+j,t,1,x).$$

[28] This relates back to our discussion of identification and exclusion restrictions in the sample selection model at the very end of Section 3. Exclusion restrictions prevent one from setting $\widetilde{g}_f(x) = g_f(x) + h(g(x))$ but shape restrictions on g and g_f can do this as well. Their "higher level assumptions" are essentially assuming that we make restrictions on g_f so that we can not add $h(g(x))$ to it and remain in the permissible class of g_f functions.

$$\lim_{\Pr(T_{i,a}>t|(X_i,Z_i)=(x,z))\to 1} E\left[Y_{i,a,a,0} \mid (X_i, Z_i) = (x, z)\right] = \mu(a, a, 0, x).$$

Thus this is a version of an "identification at infinity argument." Heckman and Navarro (2007) do not use this explicit argument because they avoid exclusion restrictions with a higher order assumption. However, they do use identification at infinity.

Identification of Φ

Next consider the identification of the cost of schooling function Φ. The best way to think about identification in these types of models is to start with the final period and work backward.

Since the maximum length of schooling is $\bar{T}$, the final decision is made when the individual has $\bar{T}-1$ years of schooling. At that point the student decides whether to attend the final year of school or not. Heckman and Navarro (2007) use an "identification at infinity" argument so that $\Pr(T_i > \bar{T}-2 \mid X_i = x, Z_i = z) \approx 1$. Then the problem becomes analogous to a static problem.[29] That is

$$\begin{aligned}
&\lim_{\Pr(T_i>\bar{T}-2|X_i=x,Z_i=z)\to 1} \Pr(T_{i\bar{T}} = \bar{T} \mid X_i = x, Z_i = z) \\
&= \Pr\Bigg(R(\bar{T}-1,\bar{T}-1,\Theta_{i,\bar{T}-1}) < \mu(\bar{T}-1,\bar{T}-1,0,x) + \varepsilon_{i,\bar{T}-1,\bar{T}-1,0} \\
&\quad - \Phi(\bar{T}-1,x,z) - \omega_{i,\bar{T}-1} + \left(\frac{1}{1+r}\right) \\
&\quad \times E\left[R(\bar{T},\bar{T},\Theta_{i\bar{T}}) \mid \Theta_{i,\bar{T}-1}\right] \mid X_i = x, Z_i = z\Bigg).
\end{aligned}$$

This is analogous to identification of the g_h function in the Roy model.[30]

Now one can just iterate backward given knowledge of all variables at $\bar{T}$ and $\bar{T}-1$. That is, the distribution of $(\frac{1}{1+r})E\left[V(\bar{T}-1,\bar{T}-1,\Theta_{i,\bar{T}-1}) \mid \Theta_{i,\bar{T}-2}\right]$ has been identified so once again we can use the identification approach of the static problem and can use the same basic style of proof. That is we can condition on a set of variables so that $\Pr(t > \bar{T}-2 \mid X_i = x, Z_i = z) \approx 1$ so that identification is analogous to the static problem. Consider the decision with $\bar{T}-2$ years of schooling.

[29] Once again, Heckman and Navarro (2007) use higher order assumptions that do not require exclusion restrictions. For example they allow for either an exclusion restriction or a cost variable to identify the scale (such as tuition described in Section 4 above).

[30] Note that we have violated one convention in this chapter which is to make conditioning explicit such as $E(\cdot \mid X_i = x)$. When we condition on $\Theta_{i,\bar{T}-1}$ we cannot do this explicitly because while the expectation inside the expression conditions on its outcome, the probability expression (immediately after the = sign) treats $\Theta_{i,\bar{T}-1}$ as a random variable.

$$\lim_{\Pr(T_i > \bar{T}-3 | X_i = x, Z_i = z) \to 1} \Pr(T_{i,\bar{T}-1} = \bar{T} - 1 \mid X_i = x, Z_i = z)$$
$$= \Pr\Bigg(R(\bar{T} - 2, \bar{T} - 2, \Theta_{i,\bar{T}-2})$$
$$< \mu(\bar{T} - 2, \bar{T} - 2, 0, x) + \varepsilon_{i,\bar{T}-2,\bar{T}-2,0} - \Phi(\bar{T} - 2, x, z)$$
$$- \omega_{i,\bar{T}-2} + \left(\frac{1}{1+r}\right) E[V(\bar{T} - 1, \bar{T} - 1, \Theta_{i,\bar{T}-1}) \mid \Theta_{i,\bar{T}-2}] \mid X_i = x, Z_i = z \Bigg).$$

One can keep iterating on this procedure so that Φ is identified in all periods.

Identification of the distribution of the error terms

Heckman and Navarro (2007) impose a factor structure so that

$$\varepsilon_{i,a,t,\ell} = \alpha'_{a,t,\ell}\tau_i + \varepsilon_{i,a,t,\ell}$$
$$\omega_{i,t} = \lambda'_t \tau_i + \xi_{i,t}$$

where τ_i is a vector random variable, the $\varepsilon's$ and $\xi's$ are all independently distributed, and the α and λ terms are factor loadings. Given this structure and that the other components of the model have been identified, identification of the distribution of the error terms and factor loadings can be done by varying the indices in much the same way as in the static model. We do not show this explicitly.

8. CONCLUSIONS

In this chapter we have presented identification results for models of the labor market. The main issue in all of these models is the issue of sample selection bias. We start with the classic Roy model and devote much space to explaining how this model can be identified. We then show how these results can be extended to more complicated cases, the generalized Roy model, treatment effect models, duration data, search models, and forward looking dynamic models. We show the importance of both exclusion restrictions and support conditions for all of these models.

TECHNICAL APPENDIX

Proof of Theorem 2.1. Let $\mathcal{X}^*$ be the set of points (x^c, x^d) at which g is continuous in x^c. For any $(x^c, x^d) \in \mathcal{X}^*$ and $\delta > 0$, $E(Y_i \mid \|X_i^c - x^c\| < \delta, X_i^d = x^d)$ is identified directly from the data.

Since g is continuous at (x^c, x^d),

$$\lim_{\delta \downarrow 0} E(Y_i \mid \|X_i^c - x^c\| < \delta, X_i^d = x^d) = g(x^c, x^d),$$

so $g(x^c, x^d)$ is identified on $\mathcal{X}^*$. By Assumption 2.2, $\mathcal{X}^*$ has measure one. □

Proof of Theorem 3.1. Let $\mathcal{X}^*$ be the set of points $(x_f^c, x_f^d, x_h^c, x_h^d, x_0^c, x_0^d)$ at which g_h and g_f are continuous in x^c.

First notice that for any $x = (x_f^c, x_f^d, x_h^c, x_h^d, x_0^c, x_0^d) \in \mathcal{X}^*$,

$$\begin{aligned}\lim_{\delta \downarrow 0} \Pr(J_i = f \mid \|X_i^c - x^c\| < \delta, X_i^d = x^d) &\equiv \Pr(J_i = f \mid X_i = x) \\ &= g(x)\end{aligned}$$

is identified.

Thus we have thus established that we can write the model as $J_i = f$ if and only if $g(X_i) > \varepsilon_i$, where ε_i is uniform $[0, 1]$ and that g is identified.

Next consider identification of g_f at the point (x_f, x_0). This is basically the standard selection problem. As long as g is continuous on the continuous covariates at this point, we can identify

$$\begin{aligned}&\lim_{\delta \downarrow 0} \text{Med}(Y_i \mid \|X_{fi}^c - x_f^c\| < \delta, X_{fi}^d = x_f^d, \|X_{0i}^c - x_0^c\| < \delta, \\ &\quad X_{0i}^d = x_0^d, |1 - g(X_i)| < \delta, J_i = f) \\ &= g_f(x_f, x_0) + \lim_{\delta \downarrow 0} \text{Med}(\varepsilon_{fi} \mid \|X_{fi}^c - x_f^c\| < \delta, X_{fi}^d = x_f^d, \|X_{0i}^c - x_0^c\| < \delta, \\ &\qquad X_{0i}^d = x_0^d, |1 - g(X_i)| < \delta, J_i = f) \\ &= g_f(x_f, x_0).\end{aligned}$$

Thus g_f is identified. Note that having an exclusion restriction with strong support conditions is necessary to guarantee that the measure of the set of X_i satisfying $|1 - g(X_i)| < \delta$ is not zero.

Next we show how to identify g_h. Note that for any (x_h, x_0) where g is continuous in the continuous covariates and $\delta > 0$ we can identify the set

$$\begin{aligned}\mathcal{X}(x_h, x_0, \delta) \equiv \{\tilde{x} \in \mathcal{X}^* : &\|\tilde{x}_h^c - x_h^c\| < \delta, \\ &\tilde{x}_h^d = x_f^d, \|\tilde{x}_0^c - x_0^c\| < \delta, \tilde{x}_{0i}^d = x_0^d, |0.5 - g(\tilde{x})| < \delta\}\end{aligned}$$

where $\tilde{x} = (\tilde{x}_f, \tilde{x}_h, \tilde{x}_0)$. Under our assumptions it has positive measure.

The median zero assumption guarantees that

$$\begin{aligned}\lim_{\delta \downarrow 0} \mathcal{X}(x_h, x_0, \delta) &= \left\{\tilde{x} \in \mathcal{X}^* : \tilde{x}_h = x_h, \tilde{x}_0 = x_0, 0.5 = \Pr(J_i = F \mid X_i = \tilde{x})\right\} \\ &= \left\{\tilde{x} \in \mathcal{X}^* : \tilde{x}_h = x_h, \tilde{x}_0 = x_0, 0.5 = \Pr(\varepsilon_{hi} - \varepsilon_{fi} \leq g_f(\tilde{x}_f, x_0) - g_h(x_h, x_0))\right\} \\ &= \left\{\tilde{x} \in \mathcal{X}^* : \tilde{x}_h = x_h, \tilde{x}_0 = x_0, g(\tilde{x}_f, x_0) = g_h(x_h, x_0)\right\}\end{aligned}$$

is identified. Since $g(\tilde{x}_f, x_0)$ is identified, g_h is identified.

Finally consider identification of G given g_f and g_h. Note that from the data one can identify

$$\begin{aligned}
&\lim_{\delta \downarrow 0} \Pr(J_i = f, \log(Y_{fi}) < s \mid \|X_i^c - x^c\| < \delta, X_i^d = x^d) \\
&\quad = \lim_{\delta \downarrow 0} \Pr(g_h(X_{hi}, X_{0i}) + \varepsilon_{hi} \leq g_f(X_{fi}, X_{0i}) + \varepsilon_{fi}, g_f(X_{fi}, X_{0i}) + \varepsilon_{fi} \\
&\qquad \leq s \mid \left\|X_i^c - x^c\right\| < \delta, X_i^d = x^d) \\
&\quad = \Pr(\varepsilon_{hi} - \varepsilon_{fi} \leq g_f(x_f, x_0) - g_h(x_h, x_0), \varepsilon_{fi} \leq s - g_f(x_f, x_0))
\end{aligned}$$

which is the cumulative distribution function of $(\varepsilon_{hi} - \varepsilon_{fi}, \varepsilon_{fi})$ evaluated at the point $(g_f(x_f, x_0) - g_h(x_h, x_0), s - g_f(x_f, x_0))$. By varying the point of evaluation one can identify the joint distribution of $(\varepsilon_{hi} - \varepsilon_{fi}, \varepsilon_{fi})$ from which one can derive the joint distribution of $(\varepsilon_{fi}, \varepsilon_{hi})$. □

Proof of Theorem 4.1. As in the proof of Theorem 3.1, let $\mathcal{X}^*$ be the set of points $(z^c, z^d, x_f^c, x_f^d, x_h^c, x_h^d, x_0^c, x_0^d)$ at which g_h, g_f, φ_h and φ_f are continuous in $(z^c, z^d, x_f^c, x_f^d, x_h^c, x_h^d, x_0^c, x_0^d)$.

First notice that for any $(z, x) = (z^c, z^d, x_f^c, x_f^d, x_h^c, x_h^d, x_0^c, x_0^d) \in \mathcal{X}^*$,

$$\begin{aligned}
&\lim_{\delta \downarrow 0} \Pr(J_i = f \mid \|X_i^c - x^c\| < \delta, \|Z_i^c - z^c\| < \delta, (Z_i^d, X_i^d) = (z^d, x^d)) \\
&\quad = \Pr(\nu_i \leq \varphi(z, x)) \\
&\quad = \varphi(z, x).
\end{aligned}$$

Thus φ is identified on the relevant set. Next consider g_f and the joint distribution of $(\nu_i, \varepsilon_{fi})$. Note that for all $(z, x_f, x_h, x_0) \in \mathcal{X}^*$ and any $y \in \mathbb{R}$, we can identify

$$\begin{aligned}
&\lim_{\delta \downarrow 0} \Pr(J_i = f, Y_{fi} \leq y \mid \|X_i^c - x^c\| < \delta, \|Z_i^c - z^c\| < \delta, (Z_i^d, X_i^d) = (z^d, x^d)) \\
&\quad = \Pr(\nu_i \leq \varphi(z, x), g_f(x_f, x_0) + \varepsilon_{fi} \leq y)
\end{aligned}$$

which is the joint distribution of $(\nu_i, g_f(x_f, x_0) + \varepsilon_{fi})$ evaluated at $(\varphi(z, x), y)$. Holding (x_f, x_0) constant and varying $(\varphi(z, x), y)$ we can estimate this joint distribution. Since the median of ε_{fi} is zero, g_f is identified and given g_f the joint distribution of $(\nu_i, \varepsilon_{fi})$ is identified. Since the model is symmetric in h and f, g_h and the joint distribution of $(\nu_i, \varepsilon_{hi})$ are identified using the analogous argument. □

Proof of Theorem 4.2. The first part is analogous to step three of identification of the Roy model presented in the text. Note that for any (z, x_0) and δ we can identify the set

$$\begin{aligned}
\mathcal{X}(z, x_0, \delta) \equiv \{(\tilde{z}, \tilde{x}) \in \mathcal{X}^* : &\|\tilde{z}^c - z^c\| < \delta, \tilde{z}^d = z^d, \|\tilde{x}_0^c - x_0^c\| < \delta, \\
&\tilde{x}_0^d = x_0^d, |0.5 - \varphi(\tilde{z}, \tilde{x})| < \delta\}
\end{aligned}$$

and it has positive measure where the elements of $(\tilde{z}, \tilde{x})$ are defined in the obvious way. The median zero assumption guarantees that

$$\begin{aligned}
&\lim_{\delta \downarrow 0} \mathcal{X}(z, x_0, \delta) \\
&\quad = \{(\tilde{z}, \tilde{x}) \in \mathcal{X}^* : \tilde{z} = z, \tilde{x}_0 = x_0, 0.5 = \Pr(J_i = F \mid (Z_i, X_i) = (\tilde{z}, \tilde{x}))\} \\
&\quad = \{(\tilde{z}, \tilde{x}) \in \mathcal{X}^* : \tilde{z} = z, \tilde{x}_0 = x_0, 0.5 = \Pr(\varepsilon_{hi} - \varepsilon_{fi} \leq g_f(\tilde{x}_f, x_0) \\
&\qquad + \varphi(z, x_0) - g_h(\tilde{x}_h, x_0)) - \varphi(z, x_0)\} \\
&\quad = \{(\tilde{z}, \tilde{x}) \in \mathcal{X}^* : \tilde{z} = z, \tilde{x}_0 = x_0, \varphi_f(z, x_0) - \varphi_h(z, x_0) \\
&\qquad = g_h(\tilde{x}_h, x_0) - g_f(\tilde{x}_f, x_0)\}.
\end{aligned}$$

Since g_h and g_f are identified by Theorem 4.1, $\varphi_f(z, x_0) - \varphi_h(z, x_0)$ is also identified. Given this we can identify the distribution of $(\varepsilon_{hi} + \nu_{hi} - \varepsilon_{fi} - \nu_{fi}, \varepsilon_{fi})$ and $(\varepsilon_{hi} + \nu_{hi} - \varepsilon_{fi} - \nu_{fi}, \varepsilon_{hi})$ since in general

$$\begin{aligned}
&\lim_{\delta \downarrow 0} \Pr(J_i = f, Y_{fi} \leq y \mid \|Z_i^c - z^c\| < \delta, Z_i^d = z^d, \|X_i^c - x^c\| < \delta, X_i^d = x_0^d) \\
&\quad = \Pr(\varepsilon_{hi} + \nu_{hi} - \varepsilon_{fi} - \nu_{fi} \leq g_f(x_f, x_0) + \varphi_f(z, x_0) - g_h(x_h, x_0) \\
&\qquad - \varphi_h(z, x_0), \varepsilon_{fi} \leq y - g_f(x_f, x_0)),
\end{aligned}$$

and

$$\begin{aligned}
&\lim_{\delta \downarrow 0} \Pr(J_i = r, Y_{hi} \leq y \mid \|Z_i^c - z^c\| < \delta, Z_i^d = z^d, \|X_i^c - x^c\| < \delta, X_i^d = x_0^d) \\
&\quad = \Pr(-\left(\varepsilon_{hi} + \nu_{hi} - \varepsilon_{fi} - \nu_{fi}\right) \leq g_h(x_h, x_0) + \varphi_h(z, x_0) - g_f(x_f, x_0) \\
&\qquad - \varphi_f(z, x_0), \varepsilon_{hi} \leq y - g_h(x_h, x_0)). \quad \square
\end{aligned}$$

Proof of Theorem 5.1. Theorem 4.1 shows that the marginal distributions of ε_{fi} and ε_{hi} are identified. Since their expectations are finite, $E(\varepsilon_{fi})$ and $E(\varepsilon_{hi})$ are identified. We also showed that g_f and g_h are identified over a set of measure 1. Note that $E(\pi_i) = E(Y_{fi}) - E(Y_{hi}) = E(g_f(X_{fi}, X_{0i}) + \varepsilon_{fi}) - E(g_h(X_{hi}, X_{0i}) + \varepsilon_{hi}) = g_f(X_{fi}, X_{0i}) - g_h(X_{hi}, X_{0i}) + E(\varepsilon_{fi}) - E(\varepsilon_{hi})$. Because all the components of $E(\pi_i)$ are identified, $E(\pi_i)$ is identified as well. $\square$

Proof of Theorem 5.2. The marginal distribution of X_i, the joint distribution of (X_i, Y_{fi}) conditional on $J_i = f$ and the joint distribution of (X_i, Y_{hi}) conditional on $J_i = h$ are identified directly from the data. Assumption 5.2 guarantees that for both fishing and hunting ($j \in \{f, h\}$), the conditional distribution of Y_{ji} conditional on X_i and $J_i = j$ is the same as the conditional distribution of Y_{ji} conditional on X_i alone. From each of these conditional distributions and the marginal distribution of X_i, one can identify $E(Y_{ji})$, and thus the average treatment effect is identified by taking the difference between the two. $\square$

Proof of Theorem 6.2. Let $\mathcal{X}^*$ be the set of points (x^c, x^d) at which the functions are all continuous in x^c.

First note that in this model the hazard rate of finding for any individual can be written as

$$\phi(X_{\lambda i}, X_{0i})\nu_i[1 - F_{i\varepsilon_f}(\log(W_i^r) - g_f(X_{fi}, X_{0i}))].$$

Our first goal is for any $(x_f, x_{\lambda,} x_0) \in \mathcal{X}^*$, to identify the values of x_h that send $g_h(x_h, x_0)$ arbitrarily large so that all offers are accepted. Since the reservation wage is strictly decreasing in g_h, the hazard rate is strictly increasing in g_h, we can do this by fixing (X_{fi}, X_{0i}) within some neighborhood of (x_f, x_0) and finding the value of x_h that minimizes the job finding rate.

More formally for any (x_f, x_λ, x_0) and δ,define

$$x_h(\delta) \equiv \operatorname{argmin} E(T_i \mid \|X_i^c - (x_f^c, x_h^c(\delta), x_\lambda^c, x_0^c)\| < \delta, X_i^d = (x_f^d, x_h^d(\delta), x_\lambda^d, x_0^d)).$$

Note that this minimum will be such that as $\delta \to 0$, $W_i^r \to 0$ so that

$$\begin{aligned}\lim_{\delta \downarrow 0} &\Pr(\log(T_i) < t, \log(W_{fit}) < w \mid \|X_i^c - (x_f^c, x_h^c(\delta), x_\lambda^c, x_0^c)\| < \delta, X_i^d \\ &= (x_f^d, x_h^d(\delta), x_\lambda^d, x_0^d)) = G_{\omega^*,\varepsilon}(t + \log(\phi(x_\lambda, x_0)), w - g_f(x_f, x_0))\end{aligned}$$

where G is the joint distribution between a convolution of ω_{it} and an extreme value and of ε_{fit}. Given G, applying the identification arguments for the mixed proportional hazard model one can identify ϕ. Furthermore, g_f can be identified through the standard argument for identification of the regression model.

Finally, recovering g_h can be done in an analogous way as for the Roy model. Notice that the reservation wage is scalable so that if we increase both C_i and W_{it} by 10%, then the reservation wage increases by 10% and the probability of job acceptance does not change. That is for any $\delta > 0$ if w_i^* solves

$$\mathrm{e}^{g_h(X_{hi}, X_{0i}) + \varepsilon_{hi}} + w_i^* = \frac{\lambda_i}{r} \int_{\log(w_i^*) - g_f(X_{fi}, X_{0i})}^{\infty} (\mathrm{e}^{g_f(X_{fi}, X_{0i}) + \varepsilon_{fit}} - w_i^*) \mathrm{d}F_{i\varepsilon_f}(\varepsilon_{fit})$$

then $w_i^* \mathrm{e}^\delta$ solves

$$\begin{aligned}&\mathrm{e}^{g_h(X_{hi}, X_{0i}) + \delta + \varepsilon_{hi}} + w_i^* \mathrm{e}^\delta \\ &\quad = \frac{\lambda_i}{r} \int_{\log(w_i^*) - g_f(X_{fi}, X_{0i})}^{\infty} (\mathrm{e}^{g_f(X_{fi}, X_{0i}) + \delta + \varepsilon_{fit}} - w_i^* \mathrm{e}^\delta) \mathrm{d}F_{i\varepsilon_f}(\varepsilon_{fit}),\end{aligned}$$

but the probability of accepting a job and thus the expected duration remains the same.

Thus as in the identification of the slope that we discuss in Step 2 of the identification of the Roy model, for any (x_h, x_0) and $(\tilde{x}_h, \tilde{x}_0)$ suppose we want to identify

$g_h(x_h, x_0) - g_h(\tilde{x}_h, \tilde{x}_0)$. Fix x_λ and $\tilde{x}_\lambda$ so that $\phi(x_\lambda, x_0) = \phi(\tilde{x}_\lambda, \tilde{x}_0)$. Then the key here is finding values x_f and $\tilde{x}_f$ so that

$$\lim_{\delta \downarrow 0} E(\log(Z(T_i)) \mid \|X_i^c - x^c\| < \delta, X_i^d = x^d)$$
$$= \lim_{\delta \downarrow 0} E(\log(Z(T_i)) \mid \|X_i^c - \tilde{x}^c\| < \delta, X_i^d = \tilde{x}^d).$$

But if this is the case it must be that

$$g_f(x_f, x_0) - g_h(x_h, x_0) = g_f(\tilde{x}_f, \tilde{x}_0) - g_h(\tilde{x}_h, \tilde{x}_0)$$

but then

$$g_h(x_h, x_0) - g_h(\tilde{x}_h, \tilde{x}_0) = g_f(x_f, x_0) - g_f(\tilde{x}_f, \tilde{x}_0)$$

where the right hand side has already been identified. Thus g_h is identified up to location on the set $\mathcal{X}^*$. □

REFERENCES

Abbring, J., 2010. Identification of dynamic discrete choice models. Annual Review of Economics 2, 367–394.

Abbring, J., Ridder, G., 2009. A note on the non-parametric identification of generalized accelerated failure-time models. Unpublished manuscript, Tilburg University.

Abbring, J., Heckman, J., 2007. Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation. In: Heckman, Leamer (Eds.), Handbook of Econometrics. North Holland, Amsterdam, pp. 5145–5303.

Abbring, J., van den Berg, G., 2003. The identifiability of the mixed proportional hazards competing risks model. Journal of the Royal Statistical Society, Series B (Statistical Methodology) 3, 701–710.

Aguirregabiria, V., 2010. Another look at the identification of dynamic discrete decision processes: an application to retirement behavior. Journal of Business and Economic Statistics 28, 201–218.

Altonji, J., Elder, T., Taber, C., 2005a. Selection on observed and unobserved variables: assessing the effectiveness of Catholic schools. Journal of Political Economy 113.

Altonji, J., Elder, T., Taber, C., 2005b. An evaluation of instrumental variable strategies for estimating the effects of Catholic schooling. Journal of Human Resources.

Angrist, J., Imbens, G., 1999. Comment on James J. Heckman, "Instrumental variables: a study of implicit behavioral assumptions used in making program evaluations". The Journal of Human Resources 34 (4), 823–827.

Angrist, J., Imbens, G., Rubin, D., 1996. Identification of causal effects using instrumental variables. Journal of the American Statistical Association 91 (June).

Angrist, J., Pischke, S., 2010, The credibility revolution in empirical economics: how better research design is taking the con out of econometrics. NBER working paper 15794.

Barlevy, G., 2008. Identification of search models using record statistics. Review of Economic Studies 75 (1), 29–64.

Bloom, H., Orr, L., Bell, S., Cave, G., Doolittle, F., Lin, W., Bos, J.s, 1997. The benefits and costs of JTPA title II-A programs: key findings from the national job training partnership act study. Journal of Human Resources 32 (3), 549–576.

Blundell, R., Gosling, A., Ichimura, H., Meghir, C., 2007. Changes in the distribution of male and female wages accounting for employment composition using bounds. Econometrica 75, 323–363.

Buera, F., 2006. Non-parametric identification and testable implications of the Roy model. Unpublished manuscript, UCLA.

Burdett, K., Mortensen, D.T., 1998. Wage differentials, employer size and labor market equilibrium. International Economic Review 39, 257–273.
Canals-Cerda, J., 2010. Identification in empirical search models when ages are measured with errors. unpublished manuscript. Federal Reserve Bank of Philadelphia.
Carneiro, P., Heckman, J., Vytlacil, E. 2010. Estimating marginal returns to education. unpublished manuscript, University College London.
Carneiro, P., Lee, S., 2009. Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality. Journal of Econometrics 149, 191–208.
Chamberlain, G., 1986. Asymptotic efficiency in semiparametric models with censoring. Journal of Econometrics 32, 189–218.
Chen, X., 2007. Large sample sieve estimation of semi-nonparametric models. In: Handbook of Econometrics. North-Holland (Chapter 76).
Das, M., Newey, W., Vella, F., 2003. Nonparametric estimation of sample selection models. The Review of Economic Studies 70 (1), 33–58.
Davidson, J., 1994. Stochastic Limit Theory. Oxford University Press, Oxford.
Deaton, A.. 2009. Instruments of development: randomization in the tropics, and the search for the elusive keys to economic development. NBER working paper 14690.
DiNardo, J., Lee, D., 2011. Program evaluation and research designs. In: Ashenfelter, Orley, Card, David (Eds.), Handbook of Labor Economics, vol. 4a. Elsevier Science, pp. 463–536.
Doyle, J., 2007. Child protection and child outcomes: measuring the effects of foster care. The American Economic Review 97 (5), 1583–1610.
Eckstein, Z., van den Berg, G, 2007. Empirical labor search: a survey. Journal of Econometrics 136, 531–564.
Elbers, C., Ridder, G., 1982. True and spurious duration dependence: the identifiability of the proportional hazard model. Review of Economic Studies 64, 403–409.
Evans, W., Schwab, R., 1995. Finishing high school and starting college: do Catholic schools make a difference?. Quarterly Journal of Economics 110, 947–974.
Flinn, C., Heckman, J., 1982. New methods for analyzing structural models of labor force dynamics. Journal of Econometrics 18, 115–168.
French, E., Song, J. 2010. The effect of disability insurance receipt on labor supply. unpublished manuscript. Federal Reserve Bank of Chicago.
Heckman, J., 1979. Sample selection bias as a specification error. Econometrica 47 (1), 153–162.
Heckman, J., 1990. Varieties of selection bias. American Economic Review 80.
Heckman, J., 1997. Instrumental variables: a study of implicit behavioral assumptions used in making program evaluations. The Journal of Human Resources 32 (3), 441–462.
Heckman, J., 1999. Instrumental variables: Response to Angrist and Imbens. The Journal of Human Resources 34 (4), 828–837.
Heckman, J., 2000. Causal parameters and policy analysis in economics: a twentieth century retrospective. Quarterly Journal of Economics 115, 45–97.
Heckman, J., Honoré, B., 1990. The empirical content of the Roy model. Econometrica 58, 1121–1149.
Heckman, J., Honoré, B., 1989. The identifiability of the competing risks model. Biometrika 76, 325–330.
Heckman, J., LaLonde, R., Smith, J., 1999. The economics and econometrics of active labor market programs. In: Ashenfelter, Card (Eds.), Handbook of Labor Economics, vol. 3A. North-Holland, New York, pp. 1865–2097.
Heckman, J., Lochner, L., Taber, C., 1998. Explaining rising wage inequality: explorations with a dynamic general equilibrium model of labor earnings with heterogeneous agents. Review of Economic Dynamics.
Heckman, J., Navarro, S., 2007. Dynamic discrete choice and dynamic treatment effects. Journal of Econometrics 136, 341–396.
Heckman, J., Robb, R., 1986. Alternative methods for evaluating the impact of interventions. In: Heckman, Singer (Eds.), Longitudinal Analysis of Labor Market Data. Cambridge University Press, New York, pp. 156–245.
Heckman, J., Singer, B., 1984a. The identifiability of the proportional hazard model. Review of Economic Studies 51, 231–241.

Heckman, J., Singer, B., 1984b. A method for minimizing the impact of distributional assumptions in econometric models for duration data. Econometrica 52, 271–320.
Heckman, J., Taber, C., 1994. Econometric mixture models and more general models for unobservables in duration analysis. Statistical Methods in Medical Research 3 (3), 279–299.
Heckman, J., Taber, C., 2008. Roy model. In: Durlauf, Blume (Eds.), The New Palgrave Dictionary of Economics Second Edition. Palgrave Macmillan.
Heckman, J., Urzúa, S., 2010. Comparing IV with structural models: what simple IV can and cannot identify. Journal of Econometrics 156 (1), 27–37.
Heckman, J., Vytlacil, E., 1999. Local instrumental variables and latent variable models for identifying and bounding treatment effects. Proceedings of the National Academy of Sciences 96, 4730–4734.
Heckman, J., Vytlacil, E., 2001. Local instrumental variables. In: Hsiao, C., Morimune, K., Powell, J. (Eds.), Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya. Cambridge University Press, Cambridge, p. 145.
Heckman, J., Vytlacil, E., 2005. Structural equations, treatment effects and econometric policy evaluation. Econometrica 73, 669–738.
Heckman, J., Vytlacil, E., 2007a. Econometric evaluation of social programs, Part I: causal models, structural models and econometric policy evaluation. In: Heckman, Leamer (Eds.), Handbook of Econometrics. North Holland, Amsterdam, pp. 4779–4874.
Heckman, J., Vytlacil, E., 2007b. Econometric evaluation of social programs, Part II: using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments. In: Heckman, Leamer (Eds.), Handbook of Econometrics. North Holland, Amsterdam, pp. 4875–5143.
Honoré, B., 1993. Identification results for duration models with multiple spells. Review of Economic Studies 60, 241–246.
Hu, Y., Shum, M., 2009. Nonparametric identification of dynamic models with unobserved state variables. Working Paper 543, Department of Economics, Johns Hopkins University.
Ichimura, H., Taber, C., 2002. Semiparametric reduced form estimation of tuition subsidies. American Economic Review 92 (2), 286–292.
Imbens, G., 2009. Better LATE than nothing: some comments on Deaton (2009) and Heckman and Urzua (2009). unpublished manuscript, Harvard University.
Imbens, G., Angrist, J., 1994. Identification and estimation of local average treatment effects. Econometrica 62.
Imbens, G., Wooldridge, J., 2009. Recent developments in the econometrics of program evaluation. Journal of Economic Literature 47 (1), 5–86.
Kashara, H., Shimotsu, K., 2009. Nonparametric identification of finite mixture models of dynamic discrete choice. Econometrica 77 (1), 135–175.
Keane, M., Wolpin, K., 2001. The effect of parental transfers and borrowing constraints on educational attainment. International Economic Review 42, 1051–1103.
Keane, M., Todd, P., Wolpin, K., 2011. The structural estimation of behavioral models: discrete choice dynamic programming methods and applications. In: Ashenfelter, Orley, Card, David (Eds.), Handbook of Labor Economics, vol. 4a. Elsevier Science, pp. 331–461.
Lalonde, R., 1986. Evaluating the econometric evaluations of training programs with experimental data. American Economic Review 76, 604–620.
Leamer, E., 1983. Let's take the con out of econometrics. American Economic Review 73 (1), 31–43.
Lee, L-F, 1978. Unionism and wage rates: a simultaneous equations model with qualitative and limited dependent variables. International Economic Review 19 (2), 415–433.
Lippman, S., McCall, J., 1976. The economics of job search: a survey, Part I. Economic Inquiry 14, 155–189.
Magnac, T., Thesmar, D., 2002. Identifying dynamic discrete decision processes. Econometrica 70 (2), 801–816.
Manski, C., 1989. Anatomy of the selection problem. The Journal of Human Resources 24 (3), 343–360.
Manski, C., 1990. Nonparametric bounds on treatment effects. American Economic Review 80 (2), 319–323.
Manski, C., 1995. Identification problems in the social sciences. Harvard University Press, Cambridge Mass..
Manski, C., 1997. Monotone treatment response. Econometrica 65 (6), 1311–1334.

Manski, C., Pepper, J., 2000. Monotone instrumental variables with an application to the returns to schooling. Econometrica 68 (4), 997–1010.
Manski, C., Pepper, J., 2009. More on monotone instrumental variables. The Econometric Journal 12 (s1), s200-s216.
Matzkin, R., 1992. Nonparametric and distribution-free estimation of the threshold crossing and binary choice model. Econometrica 60, 239–270.
Matzkin, R., 1993. Nonparametric identification and estimation of polychotomous choice models. Journal of Econometrics 58, 137–168.
Matzkin, R., 2007. Nonparametric identification. In: Heckman, Leamer (Eds.), Handbook of Econometrics. North-Holland, Amsterdam, pp. 5145–5368.
Mortensen, D., Pissarides, C., 1994. Job creation and job destruction in the theory of unemployment. Review of Economic Studies 61, 397–415.
Neal, D., 1997. The effects of catholic secondary schooling on educational attainment. Journal of Labor Economics 15, 98–123.
Neal, D., Grogger, J., 2000. Further evidence on the effects of Catholic secondary schooling. Brookings-Wharton Papers on Urban Affairs 151–193.
Postel-Vinay, F., Robin, J.-M., 2002. Wage dispersion with worker and employer heterogeneity. Econometrica 70 (6), 2295-350.
Ridder, G., 1990. The non-parametric identification of generalized accelerated failure-time models. Review of Economic Studies 57, 167–182.
Rosenzweig, M., Wolpin, K., 2000. Natural 'natural experiments' in economics. Journal of Economic Literature 38 (4), 827–874.
Roy, A.D., 1951. Some thoughts on the distribution of earnings. Oxford Economic Papers (New Series) 3, 135–146.
Rust, J., 1994. Structural estimation of Markov decision processes. In: Engle, R., McFadden, D. (Eds.), Handbook of Econometrics, vol. 4. North Holland, Amsterdam, pp. 3082–3139.
Shaikh, A., 2010. Identification in Economics, Lecture Notes for Topics in Econometrics, http://home.uchicago.edu/~amshaikh/classes/topics_winter09.html, University of Chicago.
Sims, C., 2010. Comment on Angrist and Pischke. unpublished manuscript, Princeton University.
Taber, C., 2000. Semiparametric identification and heterogeneity in dynamic programming discrete choice models. Journal of Econometrics.
Van den Berg, G., 2001. Duration models: Specification, identification and multiple durations. In: Heckman, Leamer (Eds.), Handbook of Econometrics vol. 5. Elsevier.
Vytlacil, E., 2002. Independence, monotonicity, and latent index models: an equivalence result. Econometrica 70, 331–341.
Willis, R., Rosen, S., 1979. Education and self- selection. Journal of Political Economy 87.

This page intentionally left blank

CHAPTER 7

Search in Macroeconomic Models of the Labor Market☆

Richard Rogerson*, **Robert Shimer****

* Arizona State University, United States

** University of Chicago, United States

Contents

☆ We are grateful for comments from Steven Davis, Robert Hall, Dale Mortensen, Christopher Pissarides, and the editors, for research assistance from Chris Herrington, and for financial support from the National Science Foundation.

Handbook of Labor Economics, Volume 4a
ISSN 0169-7218, DOI 10.1016/S0169-7218(11)00413-8

Abstract

This chapter assesses how models with search frictions have shaped our understanding of aggregate labor market outcomes in two contexts: business cycle fluctuations and long-run (trend) changes. We first consolidate data on aggregate labor market outcomes for a large set of OECD countries. We then ask how models with search improve our understanding of these data. Our results are mixed. Search models are useful for interpreting the behavior of some additional data series, but search frictions per se do not seem to improve our understanding of movements in total hours at either business cycle frequencies or in the long-run. Still, models with search seem promising as a framework for understanding how different wage setting processes affect aggregate labor market outcomes.

JEL classification: E24; E32; J21; J64

Keywords: Search; Unemployment; Worker flows; Business cycle; Cross-country comparisons; Wages

In the last two decades, macroeconomists have increasingly used search theory to model the labor market. The macro-search literature is now sufficiently developed to make it meaningful to assess how integrating search theory into otherwise standard aggregate models affects the analysis of macroeconomic outcomes.[1] Although search models have been used to address a wide variety of macroeconomic issues, we focus on two: short-run (business cycle) and long-run (trend) changes in aggregate labor market outcomes.

There is a hierarchy of ways in which search may be important for macroeconomic models:

1. New data: Search models draw our attention to empirical regularities and new data sets that we would typically ignore in a model without search frictions. One example is unemployment. In a search model, we can define unemployment in a manner that conforms with statistical agencies' usage: a worker is unemployed if she is not working, available for work, and actively seeking work. Models without search can at best hope to distinguish employment from non-employment, but are silent on the distinction between unemployed and inactivity (out of the labor force). A second example is worker flows. Search models make predictions about the movement of workers between employment, unemployment, and inactivity, and between employers. They can therefore be used to understand the great variety of empirical facts about job and worker flows that economists have developed during the last two decades.[2]

[1] We do not intend this chapter to be a comprehensive survey, but rather a critical assessment of the state of the literature. The standard treatment of search models remains Pissarides (2000). See also Mortensen and Pissarides (1999a,b) and Rogerson et al. (2005) for recent surveys of the labor-search literature.

[2] Abowd and Kramarz (1999) and Davis and Haltiwanger (1999) reviewed these facts in the previous volume of this handbook.

2. Search behavior: Search itself may play a special role in understanding the behavior of some aspects of the economy that are routinely studied in models without search, including total hours, employment, and wages. For example, employment may be low in some circumstances because employed workers are losing their jobs at a high rate. Alternatively, it may be low because unemployed workers are either not searching very intensively or are adopting very high reservation wages. Neither of these possibilities is easily explored in a model without search frictions. Search may also lead to new shocks, act to somehow amplify the effect of a given set of shocks to the economy, or, because of the slow adjustment of employment, propagate shocks through time.
3. Match-specific rents: Search models naturally give rise to match-specific rents.[3] This in turn implies that, even if workers and firms exploit all the bilateral gains from trade, wages are not uniquely determined by competitive forces. This richer set of possibilities for wage setting may be important in accounting for the behavior of total hours, employment, and wages. Although search is not the only source of match-specific rents, there is something distinctive about search: rents exist at the initial meeting of the worker and the firm, and so cannot be contracted away. Other mechanisms that generate match-specific rents, such as match-specific human capital and private information, only produce rents after the match has been formed.

In this chapter, we first consolidate a variety of evidence on both short run and long changes in labor market aggregates and labor market flows in the United States and other OECD countries. We then assess the ability of search models to explain this evidence. In both contexts, it is clear that search models are useful for understanding empirical regularities in unemployment and labor market flows. Beyond this, our assessment is mixed.

In the business cycle context, we argue that the existence of search frictions does not directly improve our ability to account for aggregate labor market outcomes. Three long-standing issues in business cycle research concern the amplitude, persistence, and source of fluctuations in hours and employment. Consider first the question of why employment is so volatile. Search seems a promising avenue for answering this because, at least in the United States, most of the fluctuations in employment at business cycle frequencies come from workers moving between employment and unemployment. Indeed, an increase in unemployment is often seen as the hallmark of a recession, while cyclical movements in and out of the labor force are comparatively small. But we find that, relative to a frictionless version of the real business cycle model with indivisible labor (Hansen, 1985), the presence of search frictions actually moderates the optimal extent of fluctuations in employment. In our baseline calibration this dampening effect is quantitatively significant, so that even though search does not help resolve the puzzle of why employment is so volatile, models without search are potentially abstracting from a quantitatively important factor.

[3] See the chapter by Manning in this volume for a discussion of other sources of match-specific rents.

Intuitively, search frictions act like an adjustment cost on labor and so firms fire fewer workers during downturns to avoid costly rehiring during the subsequent boom. Because search acts as an adjustment cost, it is intuitive that it serves to increase persistence. However, we find this increase to be quantitatively unimportant. Regarding the issue of what shocks cause business cycles, search models naturally give rise to an additional source of shocks relative to frictionless models: shocks to match separation probabilities. In this view, recessions might result from shocks that cause lots of existing matches to break up. Nonetheless, we find little scope for these types of shocks to account for a large share of employment fluctuations, at least in the United States.

On the other hand, recent research suggests that there is substantial scope for search models to improve our understanding of business cycle fluctuations by providing a framework for the analysis of alternative wage determination processes. Whereas the solution to a social planner's problem in a search framework does not seem to resolve any issues regarding business cycle fluctuations in the labor market, decentralized search models with different wage setting rules can improve upon their frictionless counterparts. We conclude that in the business cycle context, the main substantive contribution of search models relies on the presence of match specific rents and the opportunity for a richer set of wage setting processes.

Our analysis of long-run changes in labor market outcomes similarly leads to mixed conclusions. First, while there are substantial trend changes in relative unemployment levels across countries over time, they are still small compared with the long run decrease in hours per worker and the increase in labor force participation. From this we conclude that search frictions are unlikely to be of first-order importance in understanding long-run changes in total hours of work.

However, even if changes in unemployment are not a dominant source of differences in total hours, search theory may still help us understand these changes. A key feature of the data is that countries exhibit very different flows into and out of unemployment, even when unemployment rates are the same. Search theory is useful for assessing the role of various factors that account for these differences and how they assist us in understanding why unemployment rates have changed over time.

The direct role of search in this context remains somewhat unclear. Some research attributes an important role to how workers change their search intensity and reservation wage in response to various changes in the economic environment. But similar to recent work on business cycle fluctuations, other research attributes the most important role to how wages respond to changes in the economic environment. In these models the key role of search is to give rise to match specific rents and permit a richer set of wage responses.

An outline of the chapter follows. Section 1 focuses on business cycle fluctuations. It begins by summarizing key business cycle facts regarding total hours, employment, unemployment, and worker flows. While the emphasis is on data for the United States,

we also report comparable statistics for a range of OECD economies where available. We then present a benchmark business cycle model with search frictions and assess the ability of this model to account for the key facts relative to the frictionless equivalent. To make the models comparable, in both cases we focus on a social planner's solution, which can be decentralized through a particular assumption on wage setting. We show that search frictions per se do not improve the fit between model and data. The section closes by describing recent work which considers alternative wage-setting assumptions and has been better able to account for the business cycle facts.

Section 2 focuses on long-run changes in labor market outcomes. It begins by documenting trend changes in unemployment for a large set of OECD economies. It then compares these evolutions with evolutions for total hours, participation, hours per worker, and worker flows. We then summarize the literature that has developed to help explain the variation in long-run unemployment changes across countries, and describe in detail two recent models that feature search and provide explanations for these evolutions.

Section 3 concludes by summarizing our key findings and describing what we see as some open questions surrounding the role of search in macroeconomics. We also briefly mention more microeconomic issues in the labor market, such as the evolution of individual workers' wages and employment, where search has proven fruitful.

1. CYCLICAL FLUCTUATIONS

This section explores the ability of search models to explain the behavior of labor markets at business cycle frequencies. We break our analysis into three parts, mirroring the three ways that search may be important for macroeconomic models. First we update and extend labor market facts in an earlier volume of this *Handbook* (Lilien and Hall, 1986), highlighting the connection between those facts and the structure of search-and-matching models. We argue that search models offer a useful framework for understanding data sets that we would typically ignore in a model without frictions. We then introduce a model that integrates search theory and standard business cycle theory. To keep the comparison clean, we first focus on the solution to a social planner's problem, and so initially abstract from alternative assumptions on wage setting. We find that search itself does not resolve important puzzles in business cycle analysis, including the nature of shocks and their amplification and propagation. The final part of this section summarizes recent developments that emphasize wage rigidities in search models, a possibility that naturally arises due to the match-specific rents. It appears that such models may be important for accounting for the standard business cycle puzzles.

1.1. Facts

We begin our analysis by confirming that in the United States and most other OECD countries, most cyclical fluctuations in hours worked are accounted for by changes in

the employment-population ratio. Moreover, especially in the United States, the labor force participation rate is nearly constant, so that cyclical decreases in employment are associated with roughly equal increases in unemployment. We then consider the gross inflow and outflow of workers from unemployment, showing that recessions are characterized by a sharp spike in the inflow rate and a larger and much more persistent decline in the outflow rate. We also show that the fraction of employed workers switching jobs is procyclical, so recessions, and the initial recovery from them, are characterized by an economic environment in which it is hard to find a job. This is consistent with an aggregate matching function, where the probability that a worker finds a new job is increasing in the aggregate vacancy-unemployment ratio. Finally, we document the existence and countercyclicality of the labor wedge, a wedge between the marginal rate of substitution of consumption and leisure and the marginal product of labor. All of this evidence suggests that recessions and the early stages of recoveries are periods when workers' labor supply is constrained by the difficulty of finding a job. Search theory offers a natural framework for understanding why this may happen.

1.1.1. Hours, employment, and unemployment

Lilien and Hall (1986) decompose fluctuations in total hours worked into changes in employment and changes in hours worked per employed worker. They dismissed the relevance of search theory for explaining fluctuations in employment in the United States in part because "it has nothing to say about the shift of labor resources from employment to non-market activities that is an important part of the cycle." (Lilien and Hall, 1986, p. 1032) We update their study using data from 1976Q3 to 2009Q3 for a more comprehensive measure of hours. In contrast to the earlier chapter, we find that movements in and out of the labor force are relatively unimportant at business cycle frequencies in the United States. But in some other OECD countries, we find that fluctuations in hours per worker and movements in and out of the labor force play an important part of changes in total hours at business cycle frequencies.

For the United States, we use a quarterly series for hours worked per person aged 16 to 64 (hereafter total hours) and for the fraction of people at work, constructed following the procedure described in Cociuba et al. (2009).[4] To focus on cyclical fluctuations, we detrend the logged data using a Hodrick-Prescott (HP) filter with the usual smoothing

[4] The Bureau of Labor Statistics (BLS) constructs the underlying data series from the Current Population Survey (CPS). We first construct a monthly series for total hours, defined as the number of people at work (CPS series LNU02005053) times average hours per person at work (LNU02005054) divided by the population aged 16–64 (difference between LNU00000000 and LNU00000097), all available online since 1976Q3. We construct the monthly series for the fraction of people at work analogously. We seasonally adjust the monthly data using the Census X11 algorithm and then take quarterly averages. Following Cociuba et al. (2009), if in one month the measure of total hours is less than 96 percent of that quarter's average, we drop the month and average the remaining two months. We need this correction because the CPS measures hours worked during a reference week, the week including the 12th day of the month. As a result, measured hours worked are low in September during years when the Labor Day holiday falls on Monday, September 7: 1981, 1987, 1992, 1998, and 2009.

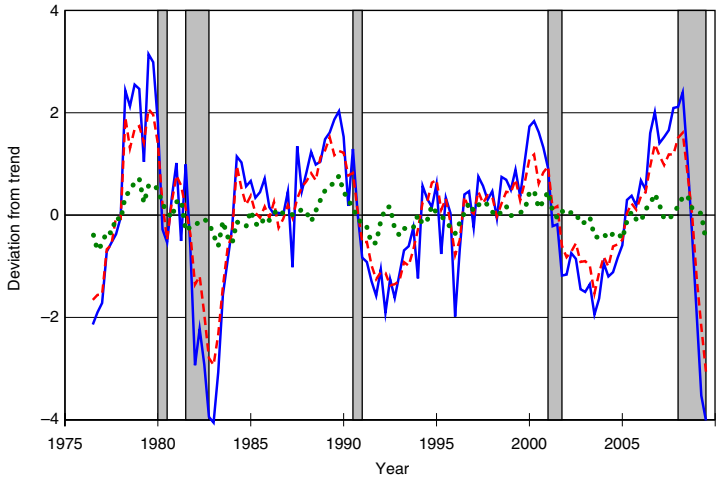


Figure 1 ***Solid line shows total hours. Dashed line shows the fraction of people at work. Dotted line shows the fraction of people in the labor force. Gray bands indicate NBER recession dates.***

parameter 1600. Cociuba et al. (2009) extend these series back to 1959 using data that are not available online. Since the rest of our data series are not available before 1976, we restrict attention to this shorter time period here.

The solid and dashed lines in Fig. 1 show the strong comovement between detrended hours and employment. The standard deviation of detrended total hours is 1.5 percent, while the standard deviation of the fraction of people at work is 1.0 percent and the correlation between the two series is 0.96. We thus conclude that, as was the case in the earlier period, "the biggest component of the variation in hours is fluctuations in the level of employment." (Lilien and Hall, 1986, p. 1006).

On the other hand, there is little change in the size of the labor force at business cycle frequencies, as shown by the dotted line in Fig. 1.[5] The standard deviation of the detrended labor force participation rate is 0.3 percent and the correlation with total hours is 0.67. For example, during the worst year of the 2008-2009 recession, from August 2008 until August 2009, total hours fell by 7.5 log points, the fraction of the population at work fell by 4.9 log points, while the size of the labor force fell by only 0.9 log points.[6] Most of the decline in total hours thus came from a decrease in employment, which was associated with a roughly equal increase in the unemployment rate. In contrast to the

[5] We measure the labor force participation rate as the number of employed people (LNU02000000) plus unemployed people (LNU03000000) divided by the population aged 16-64. We seasonally adjust and detrend the data in the same way.

[6] Recall from footnote 4 that hours data from September 2009 are low because of the timing of the Labor Day holiday. From September 2008 to September 2009, measured total hours fell by 13.2 log points, employment fell by 5.5 log points, and labor force participation by 1.2 log points.

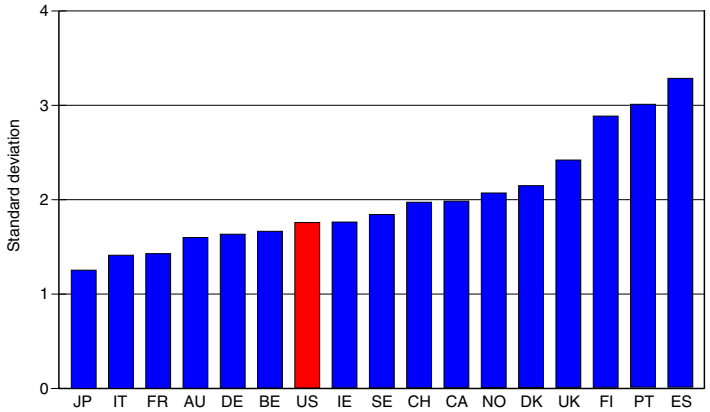


Figure 2 ***The standard deviation of detrended total hours in log points, for 17 OECD countries from 1965 to 2008.***

conclusions of Lilien and Hall (1986), recent data show that the size of the labor force is a secondary factor at business cycle frequencies in the United States.

Although the first part of this chapter is mainly focused on United States business cycle facts, we comment briefly on the extent to which these facts carry over to other advanced economies. Using OECD data on employment, unemployment, hours, and population in 17 countries from 1965 to 2008, we construct annual measures of total hours, hours per worker, the employment-population ratio, and the labor force participation rate.[7] We detrend these series using an HP filter with parameter 100, analogous to the 1600 we use elsewhere for quarterly data.

We start by looking at the standard deviation of the cyclical component of total hours. In the United States, this is 0.018 using annual data from 1965 to 2008, while the average across 17 OECD countries is slightly larger, 0.020. Figure 2 shows that there is some variation in this measure of volatility; in particular, Finland (FI), Portugal (PT), and Spain (ES) are substantially more volatile than the United States.

Figure 3 decomposes fluctuations in total hours by showing the relative standard deviation of hours per worker and the employment-population ratio. The United States is fairly typical, with a relative standard deviation 0.6, so employment accounts for most of the volatility in total hours.[8] The average across the 17 countries in our sample is even

[7] Data for civilian employment, population aged 15-64, unemployment, and unemployment durations all come from the OECD Database. Data for annual hours per worker in employment are from the Groningen Growth and Development Centre (GGDC). The countries (and country codes) we use in this paper are Australia (AU), Belgium (BE, 1965–2007), Canada (CA), Denmark (DK), Finland (FI), France (FR, 1965–2007), Germany (DE), Ireland (IE), Italy (IT), Japan (JP), Norway (NO), Portugal (PT, 1970–2008), Spain (ES), Sweden (SE), Switzerland (CH), United Kingdom (UK), and United States (US). German data are for West Germany prior to 1991 and for all of Germany starting in 1991. The sample is dictated by data availability in the OECD Database for (most of) the period since 1965.

[8] Note that the United States numbers we report here are not identical to the numbers we reported earlier. The time periods are different, the frequency of the data is different, and the underlying data sources are different.

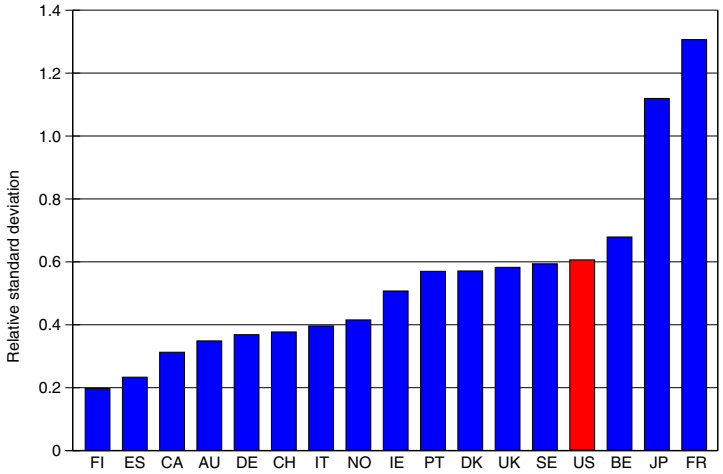


Figure 3 ***The relative standard deviation of detrended hours per worker to the detrended employment-population ratio for 17 OECD countries from 1965 to 2008.***

lower, 0.54, although this masks a significant amount of heterogeneity. For example, in France (FR) and Japan (JP), hours per worker are more volatile than the employment-population ratio, and so ignoring the hours per worker margin would seem to be a more serious omission for those countries.

Digging a bit deeper, Fig. 4 plots the correlation between total hours and each of its two components, the employment-population ratio and hours per worker. The correlation between total hours and the employment-population ratio is 0.95 in the United States, and the correlation between total hours and hours per worker is 0.86. This same pattern does not hold in all OECD countries. While the employment-population ratio is strongly correlated with total hours everywhere, with a correlation of 0.87 on average, hours per worker is not, with an average correlation 0.48. Moreover, the correlation between the employment-population ratio and hours per worker is 0.66 in the United States, while the OECD average correlation is nearly zero, 0.05. This reinforces our earlier conclusion that fluctuations in hours per worker are not a dominant source of fluctuations in total hours at business cycle frequencies in most countries.

As previously noted, France and Japan are two prominent outliers. Not only are hours per worker more volatile than the employment-population ratio, the correlation between hours per worker and total hours is higher than the correlation between the employment-population ratio and total hours. Whether this reflects institutional features of these countries, such as legislated changes in the length of the workweek in France, remains an open question.

Finally, Fig. 5 examines the cyclical component of labor force participation. Here the United States is somewhat atypical. In the United States, the correlation between

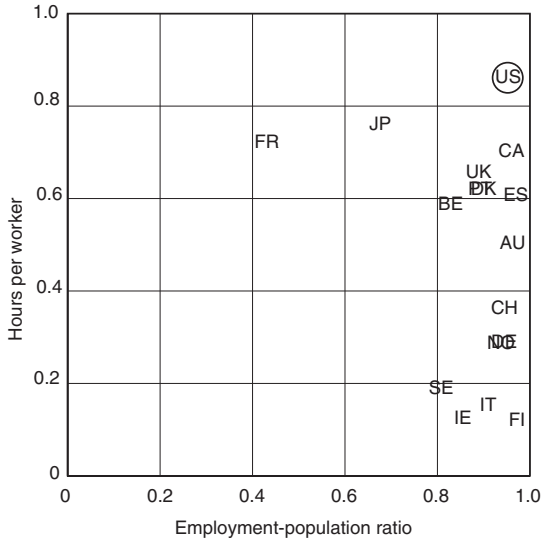


Figure 4 ***Correlation of employment-population ratio with total hours (horizontal axis) and correlation of hours per worker with total hours (vertical axis), 17 OECD countries from 1965 to 2008.***

labor force participation and the employment-population ratio is 0.68 and the standard deviation of labor force participation is 0.38 times the standard deviation of the employment-population ratio. The relative standard deviation is higher in every other country except Spain. An extreme example is Switzerland (CH), where there is essentially no difference between the standard deviation of total hours (1.80 log points), the employment-population ratio (1.84), and the labor force participation rate (1.74). Moreover, the pairwise correlation between the employment-population ratio and labor force participation rate is almost perfect. For Switzerland, most of the cyclical movement in total hours is accounted for by movements between nonparticipation and employment at a fixed number of hours per worker, and so a narrow focus only on movements between unemployment and employment would be inappropriate.

1.1.2. Unemployment inflows and outflows

Starting with Blanchard and Diamond (1990), a large literature has documented distinct cyclical patterns in the gross flows of workers between employment and unemployment. We show here that recessions are typically characterized by a sharp, short-lived increase in the inflow rate of workers from employment into unemployment and a large, prolonged decline in the outflow rate of workers from unemployment into employment. Search models are useful for making sense of these empirical regularities.

We divide our analysis of gross worker flows into several pieces. To start, we focus on the flow of workers between unemployment and employment, deferring our analysis

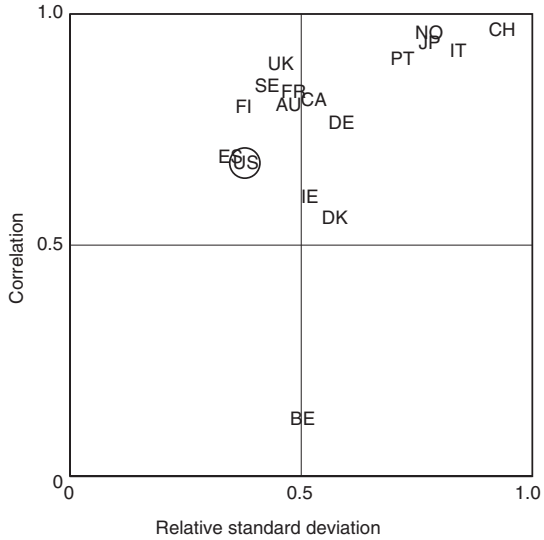


Figure 5 ***The relative standard deviation and the correlation of the detrended labor force participation rate and employment-population ratio for 17 OECD countries from 1965 to 2008.***

of inactivity (out of the labor force) until the next subsection. This abstraction enables us to construct measures of worker flows directly from publicly-available unemployment duration data. We later show that the main insights we develop here carry over to a framework that accounts for the large gross flows in and out of the labor force.

The motivation for our measurement of gross flows comes from search theory. The simplest version of the Mortensen and Pissarides (1994) search-and-matching model has a fixed labor force and a recursive structure in which the rate at which an unemployed worker finds a job, $f(t)$, depends on preferences, technology, and the state of the economy, but not directly on current unemployment $u(t)$ or employment $e(t)$. Unemployment and employment then evolve in continuous time according to

$$\dot{u}(t) = x(t)e(t) - f(t)u(t) = -\dot{e}(t),$$

where $x(t)$ is the (often exogenous) rate at which a worker exits employment for unemployment.

Our procedure for measuring the unemployment inflow rate $x(t)$ and outflow rate $f(t)$ follows Shimer (2007). Since actual data are available at discrete time intervals, once a month in the United States, we define the job finding probability F_t as the probability that a worker who starts month t unemployed finds a job within the month. Let $u_{t+1}^{<1}$ be the number unemployed with duration less than one month. Then the job finding

probability is

$$F_t = 1 - \frac{u_{t+1} - u_{t+1}^{<1}}{u_t}. \tag{1}$$

The term $u_{t+1} - u_{t+1}^{<1}$ is the number unemployed for over one month in month $t+1$. Dividing by u_t gives the fraction of the workers who failed to find a job during month t, $1 - F_t$. Assuming that the job finding rate $f(t)$ is constant during the month, Shimer (2007) proves that $F_t \equiv 1 - \mathrm{e}^{-f(t)}$, giving us the probability in the continuous time model that an unemployed worker finds at least one job during the month. Similarly, one can compute $X_t \equiv 1 - \mathrm{e}^{-x(t)}$, the probability that an employed worker loses at least one job during the month.

Using publicly available data, we can construct these series since 1948 in the United States.[9] The solid line in the top panel of Fig. 6 shows our series for the job finding probability, i.e., the outflow rate from unemployment.[10] The cyclical patterns are clear: this series rises during expansions and falls sharply during recessions. For example, during the 2008-2009 recession the job finding probability fell from 41 percent to 21 percent per month. The solid line in the bottom panel shows the employment exit probability, i.e., the inflow rate into unemployment. Here one sees a sharp, short-lived spike during most recessions. During the 2008-2009 recession, it initially rose from 2.7 to 4.3 percent per month but has since fallen back to its baseline level. The general picture is one where spikes in the unemployment inflow rate drive part of the initial increase in unemployment during most downturns, but a persistently low job finding probability explains why unemployment remains high during the subsequent recovery (Fujita and Ramey, 2009).

Elsby et al. (2008) extend this methodology to construct measures of F_t and X_t for fourteen OECD countries. They verify that the job finding probability accounts for most of the volatility in the unemployment rate in most Anglo-Saxon countries (the United States, Canada, Australia, and New Zealand), but find an equally important role for the employment exit probability in most other OECD countries, including the United Kingdom, France, Germany, and Japan. Like Fujita and Ramey (2009), they also stress

[9] The BLS constructs the underlying data series from the CPS and seasonally adjusts it using the X12 algorithm. We use data on employment (LNS12000000), unemployment (LNS13000000), and unemployment with duration 0 to 4 weeks (LNS13008396), where the latter is our proxy for the number unemployed with duration less than one month. The redesign of the CPS instrument in 1994 significantly altered the measurement of unemployment duration (Abraham and Shimer, 2001). Prior to 1994, workers were asked their unemployment duration whenever they were unemployed. After the redesign, the unemployment duration of workers who are unemployed in consecutive months is incremented by the length of the intervening time interval. To obtain a consistent series, we use the underlying microeconomic data to construct a series for short-term unemployment for workers in the "incoming rotation groups," i.e., workers who are asked about unemployment duration directly because they were not in the survey in the previous month. We seasonally adjust this data using the X11 algorithm and splice it with the official series that is available before 1994.

[10] Note that because the trends are small, we show Figs 6–10 in levels rather than detrended.

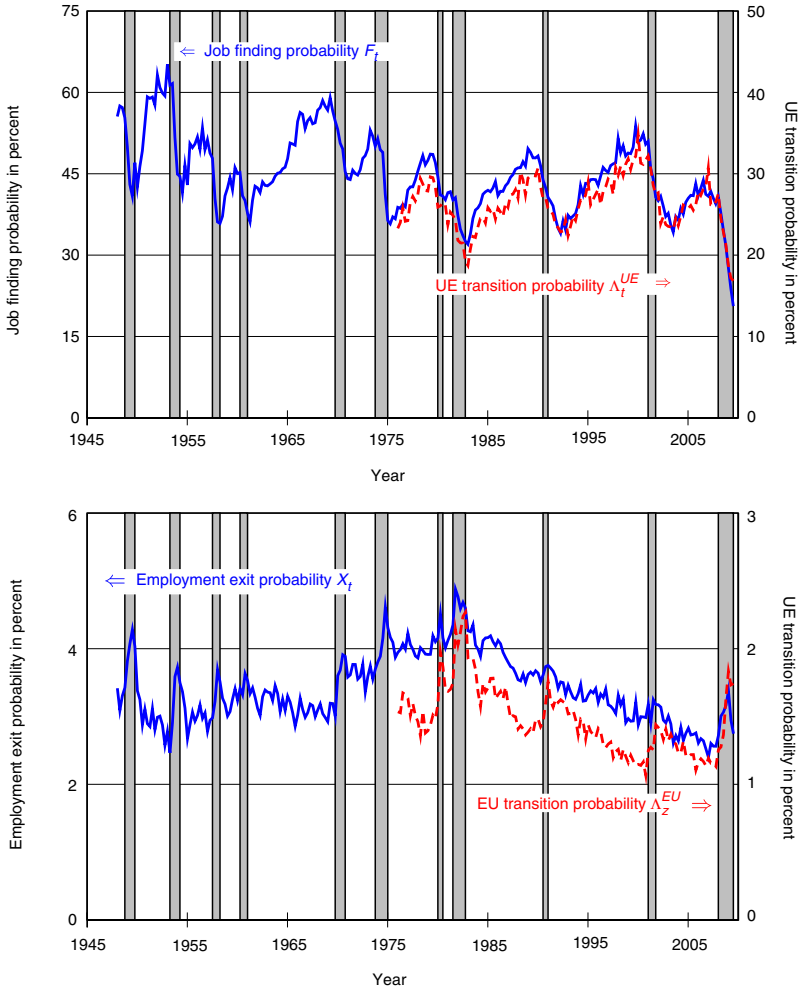


Figure 6 ***The top panel shows the job finding probability and the UE transition probability. The bottom panel shows the employment exit probability and the EU transition probability. Gray bands indicate NBER recession dates.***

that the employment exit probability is contemporaneous with the unemployment rate, while the job finding probability lags the cycle slightly.

1.1.3. Three-state model

There are interesting patterns in the flow of workers in and out of the labor force as well. During recessions, unemployed workers are not only less likely to find a job, but also less likely to drop out of the labor force. Employed workers are not only more likely to become unemployed, they are also less likely to drop out of the labor force. Similarly, inactive workers are more likely to become unemployed and less likely to find a job.

To show this, we measure gross worker flows in the United States using the monthly microeconomic data from the CPS.[11] The survey is constructed as a rotating panel, with individuals in it for four consecutive months. This means that it is theoretically possible to match up to three-quarters of the respondents between consecutive surveys, although in practice, coding errors modestly reduce the matching rate.[12] We then measure gross worker flows between labor market states A and B in month t as the number of individuals with employment status A in month $t-1$ and B in month t. This yields an updated version of the gross flows data that Blanchard and Diamond (1990) analyzed.

We manipulate this data in two ways. First, consistent with our earlier analysis, we focus on the probability that a worker switches states in a given month, rather than the total number of workers switching states—i.e., transition probabilities, rather than gross worker flows. Second we adjust the data to account for time-aggregation (Shimer, 2007). To understand why this adjustment may be important, suppose an inactive worker becomes unemployed and finds a new job within a month. We would record this as an IE transition, rather than an IU and a UI transition. Similarly, a worker may reverse an EU transition within the month, and so the job loss may disappear from the gross flows entirely. Both of these events are more likely when unemployment duration is shorter.[13]

To proceed, let λ_t^{AB} denote the Poisson arrival rate of a shock that moves a worker from state $A \in \{E, U, I\}$ to state $B \neq A$ during month t, assumed to be constant within the month. Also let $n_t^{AB}(\tau)$ denote the fraction of workers who were in state A at the start of month t and are in state B at time $t+\tau$ for $\tau \in [0,1]$. Since λ_t^{AB} is constant during the month, this satisfies an ordinary differential equation

$$\dot{n}_t^{AB}(\tau) = \sum_{C \neq B} n_t^{AC}(\tau)\lambda_t^{CB} - n_t^{AB}(\tau) \sum_{C \neq B} \lambda_t^{BC}. \tag{2}$$

11 The data since 1976 are available electronically from the National Bureau of Economic Research (NBER, http://www.nber.org/data/cps_basic.html).

12 We do not adjust the data for classification error and missing observations. Abowd and Zellner (1985) and Poterba and Summers (1986) show that misclassification in one survey creates a significant number of spurious flows. For example, Poterba and Summers (1986) show that only 74 percent of individuals who are reported as unemployed during the survey reference week in an initial interview are still counted as unemployed when they are asked in a followup interview about their employment status during the original survey reference week; 10 percent are measured as employed and 16 percent are inactive. In their pioneering study of gross worker flows, Blanchard and Diamond (1990) used Abowd and Zellner's (1985) corrected data, based on an effort by the BLS to reconcile the initial and followup interviews. Regrettably it is impossible to update this approach to the present because the BLS no longer reconciles these interviews (Frazis et al., 2005). Still, some corrections are possible. For example, the change in employment between months t and $t+1$ should in theory be equal to the difference between the flow into and out of employment. Fujita and Ramey (2009) adjust the raw gross worker flow data so as to minimize this discrepancy, as discussed in the unpublished working version of their paper. This does not substantially change the results we emphasize here.

13 Our analysis of the two-state model also implicitly accounted for time-aggregation. This is because $u_{t+1}^{<1}$ measures the share of unemployed workers with current duration less than one month, not the share of unemployed workers who were employed in the previous month. No such measure is available in the gross flows data.

$n_t^{AB}(\tau)$ increases when a worker who was in state A at t and is in state C at $t+\tau$ transitions to B and decreases when a worker who was in state A at t and is in state B at $t+\tau$ transitions to C. We solve this system of differential equations using the initial conditions $n_t^{AA}(0)=1$ and $n_t^{AB}(0)=0$ if $B \neq A$. Although the expressions are too cumbersome to include here, this gives us the six independent end-of-month shares $\{n_t^{AB}(1)\}$ as functions of the six transition rates $\{\lambda_t^{AB}\}$, where $A \neq B$. Moreover, we can measure the end-of-month shares $n_t^{AB}(1)$ directly from the gross worker flow data. To recover the instantaneous transition rates λ_t^{AB}, we invert this mapping numerically.[14]

The dashed line in the top of Fig. 6 compares the resulting three-state UE transition probability with the two-state job finding probability. The cyclical behavior of the two series is remarkably similar during the overlapping time periods, even though their levels are off by about 50 percent. The bottom panel shows that the EU transition probability likewise tracks the employment exit probability X_t, although the former is noticeably more volatile than the latter. This validates the abstraction to a two-state model for the purposes of studying United States business cycles.

Figure 7 shows transitions between all three states. Our discussion focuses on the 2008-2009 recession, although similar patterns appear in most previous recessions. The top left panel again shows the EU transition probability. From the end of 2007 to the first quarter of 2009, it rose from 1.6 to 2.3 percent per month and subsequently fell back to 2.1 percent by the third quarter of 2009. On the other hand, the top right panel indicates that employed workers were less likely to drop out of the labor force during the recession; the probability declined from 2.8 to 2.4 percent per month by the end of the sample, so the overall probability of exiting employment scarcely changed. The second row shows the probability of exiting unemployment. In this case, the decline in the UE transition probability, from 30.7 percent to 18.8 percent during the 2008-2009 recession, was reinforced by a decline in the probability of dropping out of the labor force, which fell from 28.5 to 22.1 percent. Similarly, the figure shows that inactive workers were less likely to move directly into employment and more likely to move into unemployment during the recession, further increasing the unemployment rate.

There are few comparable measures of gross worker flows in other developed economies. This appears to be a data limitation. For example, Burda and Wyplosz (1994) construct a measure of flows into and out of unemployment for France, Germany, Spain, and the United Kingdom. They use data on new registrations at unemployment offices in the first three countries and the Labor Force Survey in the United Kingdom. Thus only the last series is comparable with the methodology we describe here. They uncover significant volatility in the number of workers entering and exiting unemployment in the United Kingdom, although the movements are uncorrelated with their preferred cyclical indicator, capacity utilization. Petrongolo and Pissarides (2008) also examine dynamics

[14] If the eigenvalues of the discrete time Markov transition matrix are all positive, real, and distinct, the instantaneous transition rates are uniquely determined. In practice, this is the case in the United States.

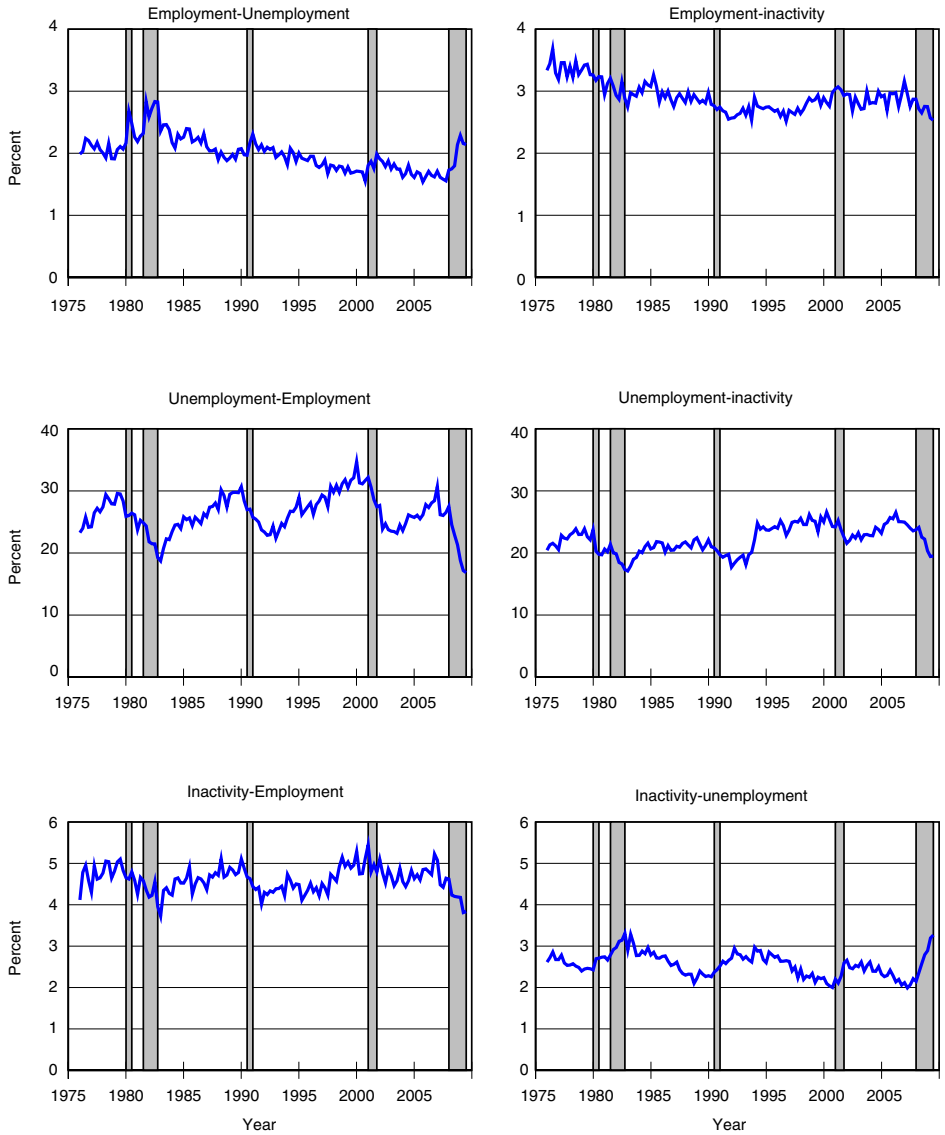


Figure 7 ***Monthly switching probabilities from state A to state B, after accounting for time aggregation (Λ_t^{AB}). Gray bands indicate NBER recession dates.***

of unemployment inflows and outflows for France, Spain and the United Kingdom, using various data sources. They find significant differences across these countries in terms of the relative importance of inflows and outflows. Ponomareva and Sheen (2009) develop a four-state model for Australia by distinguishing between part-time and full-time employment. Using data since 1981, they confirm that the job finding probability

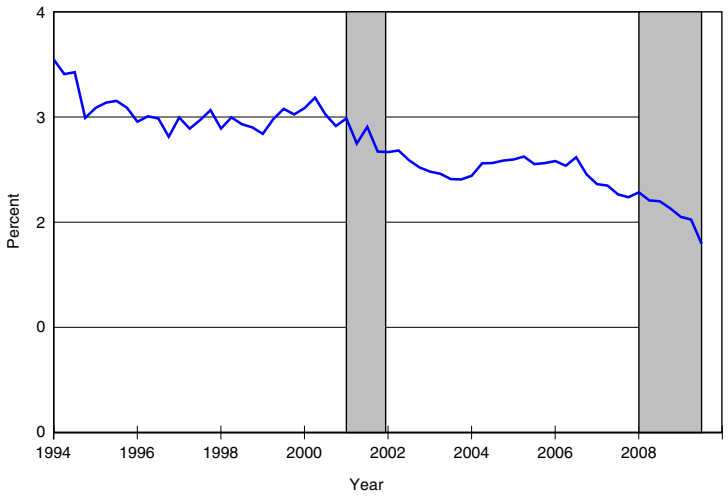


Figure 8 ***The solid line shows the fraction of workers who report switching employers during the month. Gray bands indicate NBER recession dates.***

falls sharply during recessions but also find sharp and persistent increases in the full-time employment to unemployment transition probability during recessions, particularly for men.

1.1.4. Employer-to-employer transitions

Although most job search models assume that only unemployed workers can find jobs, some newer models recognize that many workers move from employer-to-employer (EE) without experiencing an unemployment spell. Most of these papers focus on individual wage dynamics (e.g., Burdett and Mortensen, 1998; Postel-Vinay and Robin, 2002), with only a few papers examining whether EE transitions are important for understanding business cycle fluctuations (Nagypál, 2007; Moscarini and Postel-Vinay, 2008; Menzio and Shi, 2009).

Here we simply update a measure of the EE transition rate pioneered by Fallick and Fleischman (2004). Since the 1994 redesign of the CPS, respondents who are employed in consecutive months are asked "Last month, it was reported that you worked for x. Do you still work for x (at your main job)?" We use the fraction of employed workers who answer this question negatively, weighted by the CPS final weights, to compute the empirical EE transition rate. A potential shortcoming of this method is that no individual is permitted to experience multiple EE movements within a month, a possibility that may be non-negligible when the job finding rate is high. A more significant issue is that the short sample period limits any analysis of the cyclical behavior of this time series.

With these caveats, Fig. 8 shows this measure of the EE transition probability. Most noticeable is the secular decline in the rate, which is in part explained by the aging of

the United States labor force (Shimer, 2007). When it was first constructed, about 3.5 percent of workers reported switching jobs within the month, significantly higher than the EU transition probability. That number fell to 2.5 percent during the expansion from 2002 to 2007 and fell further during the subsequent recession, reaching 1.8 percent in the third quarter of 2009, somewhat below the EU transition probability. The figure suggests that the secular decline in the EE transition probability accelerates during downturns, so employed workers are less likely to switch jobs in an adverse labor market. Again, this is consistent with the evidence from gross worker flows that it is hard to find a job during downturns.

Mazumder (2007) uses data from the Survey of Income and Program Participation (SIPP) to construct a longer time series for the employer-to-employer transition probability, from 1983 to 2003. He finds a sharp increase in the series as the United States economy emerged from the 1982–1983 recession and a trough around the 1991 and 2001 recessions. This affirms that workers switch jobs at a higher rate during booms.

We are unaware of any good time series measure of the employer-to-employer transition probability outside of the United States. Fallick and Fleischman (2004, footnote 9) discuss existing studies of the extent of employer-to-employer transitions using data from other OECD countries; however, none of these papers constructs a consistent time series.

1.1.5. Matching function

We have argued that unemployment rises during recessions both because employed workers are more likely to lose their job and unemployed workers are less likely to find a job. Although an exact decomposition remains controversial, many studies suggest that the decline in the job finding probability is more important than the increase in the employment exit probability, at least in the US (Shimer, 2007; Elsby et al., 2009). In any case, regardless of the empirical evidence, search and matching models have to a large extent focused on fluctuations in the probability of finding a job. Recessions in this view are times when unemployed workers stay unemployed longer. The question is, "Why?"

Search and matching models explain fluctuations in the job finding probability through the matching function (Pissarides, 1985). The number of new matches created in month t, m_t, is a function of unemployment u_t and job vacancies v_t, say $m_t = m(u_t, v_t)$. This implies that the job finding probability for the average unemployed worker is $F_t = m(u_t, v_t)/u_t$. A common assumption is that the matching function has constant returns to scale,[15] so the probability that an unemployed worker finds a job is a function only of the vacancy-unemployment ratio, $F_t = f(\theta_t)$, where $\theta_t \equiv v_t/u_t$ is often called "market tightness."

[15] Diamond (1982a) showed how increasing returns to scale in the matching process can create multiple equilibria. The subsequent literature has found scant evidence for increasing returns, however; see the survey by Petrongolo and Pissarides (2001).

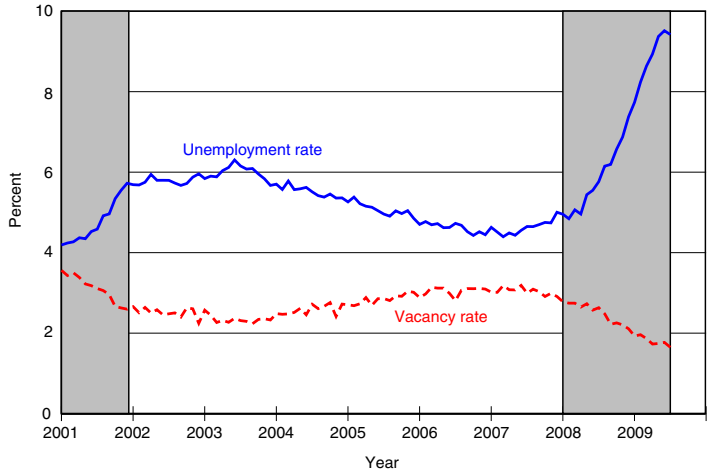


Figure 9 ***The solid line shows the unemployment rate. The dashed line shows the vacancy rate. Gray bands indicate NBER recession dates.***

At one level, we know the matching function is an incomplete description of the job finding probability because unemployed workers are not the only ones who find jobs. We have already shown that some inactive workers move directly into employment and that some employed workers switch jobs without an intervening unemployment spell. It is therefore remarkable that this simple theoretical structure describes the comovement of the job finding probability and market tightness so well.

To show this, we use data on job vacancies from the Job Openings and Labor Turnover Survey (JOLTS), a monthly survey of 16,000 business establishments.[16] According to the survey form, a job opening must satisfy three conditions: "A specific position exists; work could start within 30 days; and [the employer is] actively seeking workers from outside this location to fill the position." The survey started in December 2000, at the peak of a business cycle, and has since followed a modest expansion and strong recession. Figure 9 shows the strong negative correlation between the unemployment rate and the vacancy rate, defined as vacancies divided by vacancies plus employment.[17] This stable relationship is called the Beveridge curve.

Since unemployment is strongly negatively correlated with vacancies and with measures of the job finding probability, market tightness is strongly positively correlated with the job finding probability. Figure 10 shows the close link between a three month

16 We use BLS series JTS00000000JOL, total non-farm job openings.

17 Davis et al. (2008) argue that there are significant measurement problems in the JOLTS. These mostly show up in labor turnover statistics, but they find that job openings are unreported by about eight percent, with little cyclical variation in measurement error. The BLS has since modified the reported JOLTS data to address these concerns, but in any case, this type of error would not substantially change our conclusions.

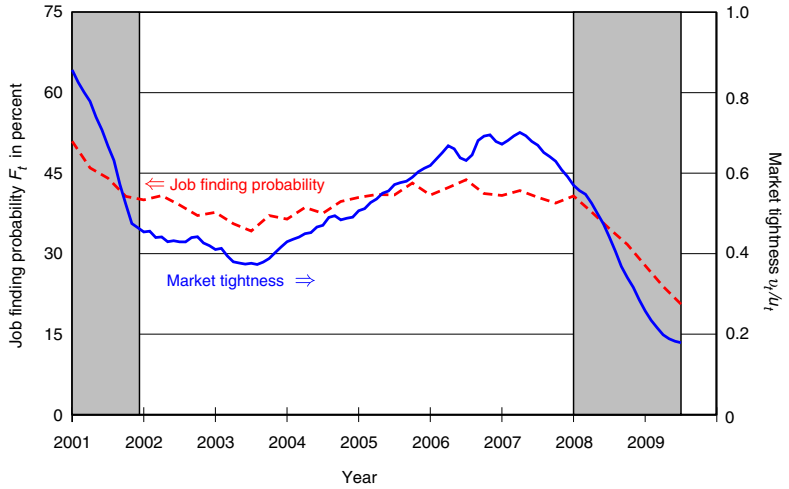


Figure 10 ***The solid line shows the vacancy-unemployment ratio. The dashed line shows the job finding probability. Gray bands indicate NBER recession dates.***

moving average of market tightness, $\theta_t = v_t/u_t$, and the quarterly series for the job finding probability. Clearly market tightness is more volatile than the job finding probability. To quantify this, suppose that the matching function is Cobb-Douglas, $m_t = \bar{m} v_t^{\eta} u_t^{1-\eta}$ for some constants $\bar{m}$ and η. Then the job finding probability should be a constant elasticity function of market tightness, $F_t = m_t/u_t = \bar{m}\theta_t^{\eta}$. We assume that there is multiplicative noise, e.g., measurement error, that disturbs this equation. Using the underlying monthly data, we then use OLS to estimate $\eta = 0.42$ with a standard error of 0.02.

An obvious shortcoming of the JOLTS is its brevity. Prior to 2001, the best available measure of job vacancies came from the Conference Board help-wanted advertising index. Abraham (1987) discusses this measure in detail, showing that it tracks job vacancies in regions where both series are available. Using data from 1951 to 2003, Shimer (2005b) estimates an elasticity of $\eta = 0.28$, somewhat smaller than the number we obtain from the shorter JOLTS data. Whether this reflects the peculiarities of help-wanted advertising, of the JOLTS data, or of the last decade remains an open question.

The survey by Petrongolo and Pissarides (2001) discusses many earlier estimates of the matching function from the United States and a variety of European countries. Most papers are interested in whether the matching function exhibits constant returns to scale and so do not constrain the coefficients on unemployment and vacancies to sum to 1. Typically they find that the coefficient on vacancies is larger than the coefficient on unemployment, in contrast to the evidence from JOLTS and the help-wanted index, although the exact estimates differ significantly across countries. And typically they cannot reject constant returns. In any case, all of these papers establish a robust,

but heterogeneous, link between unemployment, job vacancies, and the probability of finding a job. This is consistent with one of the key building blocks of search models.

1.1.6. Labor wedge

A final fact that supports the empirical relevance of search theoretic models of the labor market is evidence that workers are constrained in their ability to supply labor during recessions. One way to express this concretely is to note that, from the perspective of a labor-market-clearing model, recessions appear to be times when there is an increase in the tax on labor. A large literature has observed this fact in United States data, noted that it is hard to observe any real movements in tax rates at these frequencies, and instead called the tax a "labor wedge" (see, for example, Parkin, 1988; Rotemberg and Woodford, 1991, 1999; Hall, 1997; Mulligan, 2002; Chari et al., 2007; Shimer, 2010). The existence of a counter-cyclical labor wedge is a more nuanced assertion than the other facts we document in this chapter, since it depends on some assumptions about preferences and technology. Still, it accords with many economists' intuition that workers are not always on their labor supply curve.

We assume that a representative worker has time-separable preferences, with period utility function $u(c, h)$ defined over consumption c and hours h. We impose two restrictions on the period utility function: it must be consistent with balanced growth, so the income and substitution effects in labor supply cancel; and it must have a constant Frisch elasticity of labor supply $\varepsilon > 0$. Shimer (2010) and Trabandt and Uhlig (2009) show that these restrictions together impose the functional form

$$u(c,h) = \frac{c^{1-\sigma}(1 + \frac{\gamma(\sigma-1)\varepsilon}{\sigma(1+\varepsilon)} h^{\frac{1+\varepsilon}{\varepsilon}})^{\sigma} - 1}{1-\sigma},$$

where $\sigma > 0$ is a measure of the complementarity between hours worked and consumption, and $\gamma > 0$ is the disutility of labor supply. The restriction to balanced growth preferences is quantitatively important for our results, and so we discuss it further below. In contrast, the assumption that the labor supply elasticity ε is constant is less important for our analysis. It is useful because, as we show below, ε is a key parameter for determining the magnitude of fluctuations in the labor wedge. Note that if the complementarity parameter is fixed at $\sigma = 1$, preferences reduce to

$$u(c,h) = \log c - \frac{\gamma\varepsilon}{1+\varepsilon} h^{\frac{1+\varepsilon}{\varepsilon}},$$

additively separable between consumption and leisure. If $\sigma > 1$, the marginal utility of consumption is increasing in hours worked, creating a tendency towards a positive co-movement between consumption and labor supply.

The worker faces a period budget constraint

$$b_t = a_t + (1 - \tau_t) w_t h_t - c_t.$$

She enters a period with some initial financial wealth a_t, earns a pre-tax wage w_t per hour of work h_t, pays a proportional labor tax τ_t, and consumes c_t, leaving her with financial wealth b_t, which is then invested in any available assets. We include time subscripts on consumption, hours, and the wage to stress that these are likely time-varying. We also include a time subscript on the labor tax because our methodology will uncover cyclical fluctuations in it. We stress that this formulation is consistent with either complete or incomplete asset markets.

The key assumption is that a worker is free to increase or decrease both her consumption and labor supply. This means she can always finance an extra $(1-\tau_t)w_t$ units of consumption by working for an additional hour or she can reduce her consumption by this amount by working one hour less. In particular, a worker who maximizes lifetime utility subject to a sequence of budget constraints will set the marginal rate of substitution between consumption and leisure equal to the after-tax wage. In period t, this gives

$$\frac{\gamma c_t h_t^{\frac{1}{\varepsilon}}}{1 + \frac{\gamma(\sigma-1)\varepsilon}{\sigma(1+\varepsilon)} h_t^{\frac{1+\varepsilon}{\varepsilon}}} = w_t(1 - \tau_t). \tag{3}$$

Equation (3) gives a necessary condition from the worker's optimization problem in a variety of economic environments.

We also assume that a representative firm has access to a Cobb-Douglas production technology which uses capital k and labor h to produce output. The firm chooses its inputs to maximize its period profits

$$A_t k_t^{\alpha} h_t^{1-\alpha} - r_t k_t - w_t h_t,$$

where A_t is total factor productivity, α is the capital share of income, and r_t is the rental rate on capital. Letting $y_t = A_t k_t^{\alpha} h_t^{1-\alpha}$ denote total output, the first order condition for the choice of labor is

$$(1 - \alpha) y_t / h_t = w_t, \tag{4}$$

which equates the marginal product of labor to the wage. Again we include time subscripts on output, hours, and the wage. Note that this holds as long as the firm is free to vary its labor at a constant wage rate w_t.

Now eliminate the wage between Eqs (3) and (4). Since the hours choice of the representative household and the representative firm are equal in equilibrium, we can

write this as

$$1 - \tau_t = \frac{(\frac{\gamma}{1-\alpha})(\frac{c_t}{y_t})h_t^{\frac{1+\varepsilon}{\varepsilon}}}{1 + \frac{\gamma(\sigma-1)\varepsilon}{\sigma(1+\varepsilon)}h_t^{\frac{1+\varepsilon}{\varepsilon}}}. \tag{5}$$

The left hand side is the proportion of labor income left after taxes. The right hand side includes several different objects: the consumption-output ratio c_t/y_t, hours worked h_t raised to an exponent $(1 + \varepsilon)/\varepsilon \geq 1$, and some constants. The constants include preference parameters (σ, γ, and ε) and a technology parameter (α). Treating the constants as, in fact, constant at business cycle frequencies, the labor-market-clearing model predicts some comovement between the consumption-output ratio and hours worked in response to a shock to any variable not in this equation, such as productivity or government spending.

To explore whether this relationship is a good description of the data, we use empirical measures of the consumption-output ratio and hours worked in the United States,[18] fix the capital share at a conventional value of $\alpha = 1/3$, and consider different values for the complementarity parameter σ, the elasticity of labor supply ε, and the disutility of work γ. In all cases, we set parameters so that the average labor tax is $\tau = 0.4$.[19]

To start, we set the complementarity parameter at $\sigma = 1$, so preferences are additively separable between consumption and leisure. This case is particularly convenient because the percent deviation of $1 - \tau_t$ from trend does not depend on the choice of the parameters γ and α. Rather, the consumption-output ratio is supposed to be inversely proportional to hours worked raised to the power $(1 + \varepsilon)/\varepsilon$. In the data, we find that, while there is a negative correlation between the consumption-output ratio and hours worked, hours are more volatile than c/y. This is inconsistent with Eq. (5) for any value of the labor supply elasticity ε. Instead, recessions look like times when workers would like to supply more labor at a wage equal to the marginal product of labor, implying that it is as if the tax on labor has increased.

Figure 11 quantifies these statements. We show time series for the labor wedge for three different values for the Frisch labor supply elasticity, $\varepsilon = \frac{1}{2}$, 1, and ∞, corresponding to $(1 + \varepsilon)/\varepsilon = 3$, 2, and 1. In all cases, the implied value of $1 - \tau_t$ falls sharply during recessions, i.e., the labor wedge is countercyclical. The 2008-2009 recession stands out. Even with an infinite labor supply elasticity, $1 - \tau_t$ stood about 3 log points above trend at the beginning of 2008 and fell to 3 log points below trend

[18] We measure consumption as nominal nondurable and service consumption and output as nominal GDP from the National Income and Product Accounts, Table 1.1.5. Our measure of hours worked is unchanged from Section 1.1.1.

[19] In a more complete model, this represents a combination of income taxes, payroll taxes, and consumption taxes. This is a reasonable value for the average marginal tax rate in the United States; see Prescott (2004).

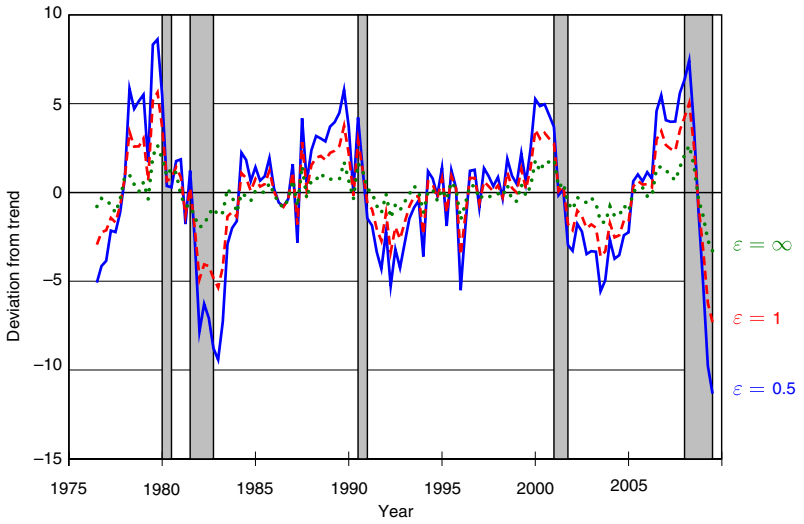


Figure 11 ***Deviation of*** $1-\tau$ ***from trend,*** $\sigma=1$***, for three different values of the Frisch labor supply elasticity*** ε***. Gray bands indicate NBER recession dates.***

in the latest available data, in the third quarter or 2009. Smaller labor supply elasticities exacerbate this issue.

Figure 12 examines the role of consumption-hours complementarity σ. We fix $\varepsilon=1$ and consider three different values, $\sigma=1$, $\sigma=2$, and $\sigma=\infty$. In all cases we fix γ so that the average value of τ is 0. Higher complementarity implies that the denominator in Eq. (5) is more sensitive to hours. In particular, when $\sigma>1$, a decline in hours lowers the denominator and so raises $1-\tau_t$, improving the fit of the model and data. But Fig. 12 shows that the improved fit is modest, even in the limit with maximum curvature.

Since the labor wedge only requires data on the consumption-output ratio and hours, it is straightforward to construct a time-series measure of it in other countries. We use data on the consumption-output ratio from the Penn World Tables and on total hours from Section 1.1.1. We assume consumption and leisure are separable, $\sigma=1$, which ensures that the disutility of work γ and the capital share α do not affect the volatility of $1-\tau_t$. Finally, we fix the elasticity of labor supply at $\varepsilon=1$.[20]

Figure 13 compares the relative standard deviation of $1-\tau_t$ and total hours with the correlation between these two series for 17 OECD countries using annual data from 1965 to 2004. Although the correlation is higher in the United States than in any other country except the United Kingdom, at 0.98, it exceeds 0.5 in every country and is below

[20] The NBER working paper version of Ohanian et al. (2008) measures both the trend and the cyclical component of the labor wedge in 21 OECD countries. They assume preferences consistent with a constant Frisch elasticity of leisure (rather than labor supply) and set this equal to −1. Their findings are broadly similar to the ones we develop here.

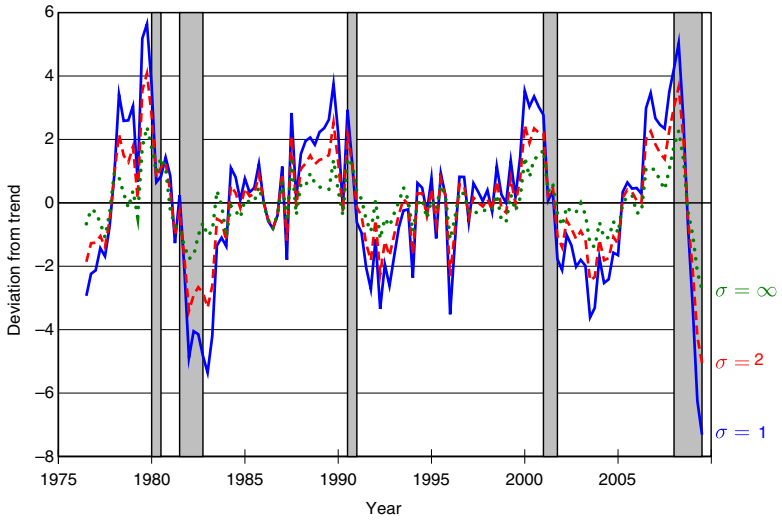


Figure 12 ***Deviation of*** $1-\tau$ ***from trend,*** $\varepsilon = 1$***, for three different values of the complementarity parameter*** σ***. Gray bands indicate NBER recession dates.***

0.75 only in Japan. Moreover, the relative volatility of the labor wedge is smaller in the United States than in most other countries. The results are reasonably robust to higher values of the labor supply elasticity. Even with $\varepsilon = \infty$, we find that the correlation between $1-\tau_t$ and h_t remains above 0.5 in every country except Japan, where it falls virtually to 0.

The bottom line is that, with these functional forms for preferences and technology, recessions look like times when the labor income tax rises. One possible interpretation is that workers are constrained from working as much as they would like during recessions, perhaps because search frictions prevent them from finding a job. Although we will explore this possibility further in the theoretical portion of this chapter, it is worth noting that with other preferences, the puzzle would disappear.[21] For example, suppose

$$u(c,h) = c - \frac{\gamma}{1+\varepsilon} h^{\frac{1+\varepsilon}{\varepsilon}}.$$

In this case, the labor wedge Eq. (5) becomes

$$1-\tau = \frac{(\frac{\gamma}{1-\alpha}) h_t^{\frac{1+\varepsilon}{\varepsilon}}}{y_t}.$$

[21] Alternatively, considering richer heterogeneous agent models may be useful in accounting for the labor wedge (Chang and Kim, 2007).

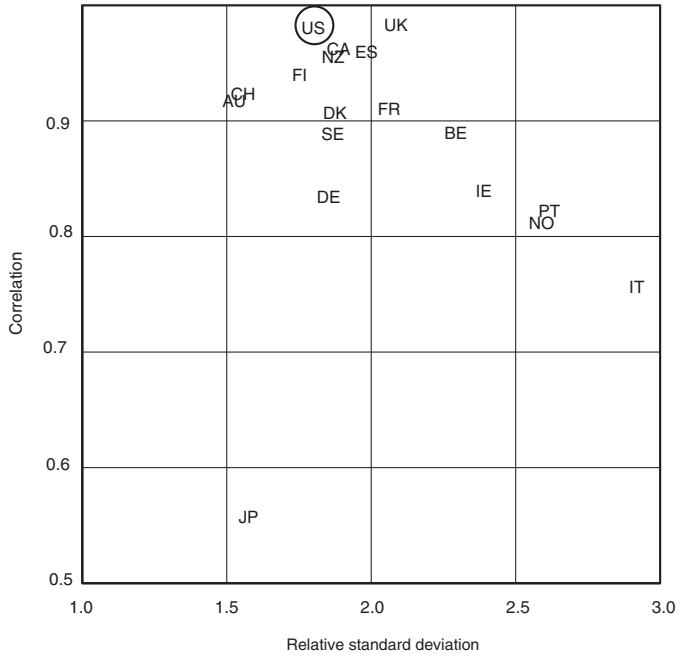


Figure 13 ***The relative standard deviation and the correlation of detrended*** $1-\tau_t$ ***and detrended total hours, for 17 OECD countries from 1965 to 2004.***

Since output is somewhat more volatile than hours and the two outcomes are strongly positively correlated, the labor wedge is not particularly cyclical when ε is sufficiently large. But we view these preferences, in particular the absence of income effects in labor supply, as implausible. They imply that at a point in time, high wage workers should supply far more labor than low wage workers. Similarly, they imply that over time, there should be a strong increasing trend in hours worked. Neither of these patterns is in the data. On the contrary, Fig. 21 suggests that hours worked may be falling in the long-run. As we discuss in the second section of this chapter, part of that may be a response to rising labor and consumption taxes, but to the extent this reflects a deviation from balanced-growth preferences, it suggests that income effects are stronger than substitution effects. In this case, the results we have reported here understate the cyclicality of the labor wedge.

1.2. Theory

This section uses an explicit dynamic stochastic equilibrium search and matching model to explore whether search frictions are useful for explaining this set of business cycle facts. Our treatment here follows Shimer (2010), which in turn builds on the canonical model

in Pissarides (1985) and early efforts to integrate that model into the real business cycle framework (Merz, 1995; Andolfatto, 1996).

1.2.1. Model setup

The model examines the interaction between a representative firm and a representative household in a closed economy. The firm uses a standard production technology combining capital and labor to create a single final good, which is used both for consumption and investment. It also uses labor to recruit more workers. At the firm level, the recruiting technology is constant returns to scale, but the efficiency of recruiting is decreasing in the aggregate recruiter-unemployment ratio.

The household has preferences over consumption and leisure that are additively separable over time and between consumption and leisure. It has many members and so can insure individuals against idiosyncratic risk, a standard device for finding a complete markets allocation. The household inelastically supplies workers to the market, although not all of them are always employed due to the search frictions.

We focus on a planner's problem, where the planner chooses consumption, investment, and the allocation of workers to production and recruiting to maximize the household's utility subject to the economy's resource constraint and the search frictions. By looking at the planner's problem, we can understand how search frictions per se affect the behavior of aggregate labor market outcomes. Section 1.3 discusses the importance of wage setting, emphasizing that "rigid wages" may arise due to match-specific rents and can significantly affect labor market outcomes.

Finally, note that our formulation abstracts from distortionary taxes, and so the labor wedge is non-zero only because search frictions create a gap between the marginal product of labor and the marginal rate of substitution between consumption and leisure.[22]

Time and states We study a discrete time model with an infinite horizon. Denote time by $t = 0, 1, 2, \ldots$ and the state of the economy at time t by s_t. Let $s^t = \{s_0, s_1, \ldots, s_t\}$ denote the history of the economy and $\Pi(s^t)$ denote the time-0 belief about the probability of observing an arbitrary history s^t through time t.[23] Aggregate productivity and the probability of exiting employment are both exogenous functions of history s^t.

Households A representative household has preferences over consumption $c(s^t)$ and labor supply $n(s^t)$ in history s^t, ordered by

$$\sum_{t=0}^{\infty} \sum_{s^t} \beta^t \Pi(s^t)(\log c(s^t) - \gamma n(s^t)). \tag{6}$$

[22] It is straightforward to introduce a distortionary labor tax with the proceeds rebated lump-sum to households. If the planner does not internalize the tax rebate, the distortionary tax creates an additional labor wedge. This does not affect the model's cyclical properties.

[23] One might instead call the probability $\Pi_t(s^t)$ to clarify that the length of the vector s^t depends on t. With a slight abuse of notation we simply call this $\Pi(s^t)$, and similarly for other history-dependent functions.

Note that labor is indivisible and each individual suffers a utility loss γ when employed. Thus the household effectively has an infinite Frisch elasticity of labor supply. Indivisible labor implies $n(s^t)$ is the employment rate and, normalizing the size of the household, $1 - n(s^t)$ is the unemployment rate.

Recruiting Let $\theta(s^t)$ denote the aggregate ratio of recruiters to unemployed workers in history s^t. Employment evolves as

$$n(s^{t+1}) = (1 - x(s^t))n(s^t) + f(\theta(s^t))(1 - n(s^t)) \tag{7}$$

where $s^{t+1} \equiv \{s^t, s_{t+1}\}$ is a continuation history of s^t. A fraction $x(s^t)$ of the employed workers lose their job and become unemployed, while the remainder stay employed. A fraction $f(\theta(s^t))$ of the $1 - n(s^t)$ unemployed workers find a job. We assume that the job finding probability f is increasing in the recruiter-unemployment ratio. Conversely, each recruiter attracts $\mu(\theta(s^t))$ workers to the firm, where $\mu(\theta) \equiv f(\theta)/\theta$; this is a decreasing function. Thus at the firm level, the recruiting technology has constant returns to scale. This formulation is inspired by the empirical evidence supporting the existence of a matching function, although we model the inputs into the matching function as unemployed workers and recruiters, rather than unemployment and vacancies.[24]

Production Firms have access to a standard Cobb-Douglas production technology. Total output is

$$y(s^t) = k(s^t)^\alpha (z(s^t)(n(s^t) - \theta(s^t)(1 - n(s^t))))^{1-\alpha},$$

where $k(s^t)$ is the capital stock, $z(s^t)$ is labor-augmenting productivity, and $\theta(s^t)$ $(1 - n(s^t))$ is the number of recruiters (the recruiter-unemployment ratio times unemployment), so $n(s^t) - \theta(s^t)(1 - n(s^t))$ is the number of workers used in production, i.e., producers. Thus the economy faces a resource constraint

$$k(s^{t+1}) = k(s^t)^\alpha (z(s^t)(n(s^t) - \theta(s^t)(1 - n(s^t))))^{1-\alpha} + (1 - \delta)k(s^t) - c(s^t), \tag{8}$$

where δ is the fraction of capital that depreciates in production.

1.2.2. Planner's problem

The planner starts history s^t with capital k and employment n. He chooses how much each individual consumes and the recruiter-unemployment ratio. Next period's

[24] See Mortensen and Pissarides (1999a,b) for excellent surveys of search models with frictions based on the matching function.

employment then follows Eq. (7), while next period's capital stock satisfies the resource constraint (8). Let $V(k, n, s^t)$ denote the expected utility of the representative household when the aggregate capital stock is k, aggregate employment is n, and the history is s^t. Expressing the planner's problem recursively gives

$$V(k, n, s^t) = \max_{c,\theta,k',n'} \log(c) - \gamma n + \beta \sum_{s^{t+1}|s^t} \frac{\Pi(s^{t+1})}{\Pi(s^t)} V(k', n', s^{t+1}),$$

where our notation assumes s^{t+1} is a continuation history of s^t. The planner recognizes that next period's employment satisfies

$$n' = (1 - x(s^t))n + f(\theta)(1 - n)$$

and next period's capital stock satisfies

$$k' = k^\alpha (z(s^t)(n - \theta(1 - n)))^{1-\alpha} + (1 - \delta)k - c.$$

To solve this, substitute the laws of motion for employment and capital into the value function and then take the first order condition for consumption:

$$\frac{1}{c(s^t)} = \beta \sum_{s^{t+1}|s^t} \frac{\Pi(s^{t+1})}{\Pi(s^t)} V_k(k(s^{t+1}), n(s^{t+1}), s^{t+1}). \tag{9}$$

The left hand side is the marginal utility of consumption, while the right hand side is the expected discounted marginal value of investment. Next take the first order condition for recruiting, assuming an interior solution

$$\begin{aligned}(1 - \alpha) z(s^t) \kappa(s^t)^\alpha \sum_{s^{t+1}|s^t} \frac{\Pi(s^{t+1})}{\Pi(s^t)} V_k(k(s^{t+1}), n(s^{t+1}), s^{t+1}) \\ = f'(\theta(s^t)) \sum_{s^{t+1}|s^t} \frac{\Pi(s^{t+1})}{\Pi(s^t)} V_n(k(s^{t+1}), n(s^{t+1}), s^{t+1}),\end{aligned} \tag{10}$$

where

$$\kappa(s^t) \equiv \frac{k(s^t)}{z(s^t)(n(s^t) - \theta(s^t)(1 - n(s^t)))} \tag{11}$$

is the equilibrium capital-labor ratio in the production sector. The left hand side of Eq. (10) is the marginal product of labor, multiplied by the value of the capital lost by a

small increase in the recruiter-unemployment ratio. The right hand side is the increase in employment from the shift in the ratio, multiplied by the marginal value of employment.

Next turn to the envelope condition for capital:

$$V_k(k(s^t), n(s^t), s^t) = (\alpha\kappa(s^t)^{\alpha-1} + 1 - \delta)\beta \sum_{s^{t+1}|s^t} \frac{\Pi(s^{t+1})}{\Pi(s^t)} V_k(k(s^{t+1}), n(s^{t+1}), s^{t+1}). \quad (12)$$

The marginal value of capital today is the gross marginal product of capital net of depreciation multiplied by the expected marginal value of capital tomorrow. Finally, the envelope condition for labor:

$$\begin{aligned} V_n(k(s^t), n(s^t), s^t) &= -\gamma + (1-\alpha)z(s^t)\kappa(s^t)^{\alpha} \\ &\times (1 + \theta(s^t))\beta \sum_{s^{t+1}|s^t} \frac{\Pi(s^{t+1})}{\Pi(s^t)} V_k(k(s^{t+1}), n(s^{t+1}), s^{t+1}) \\ &+ (1 - x(s^t) - f(\theta(s^t)))\beta \sum_{s^{t+1}|s^t} \frac{\Pi(s^{t+1})}{\Pi(s^t)} V_n(k(s^{t+1}), n(s^{t+1}), s^{t+1}). \quad (13) \end{aligned}$$

The marginal value of employment is the disutility of work, plus the value of the increase in next period's capital stock that comes from the additional output, plus the value of the increase in next period's employment that comes from having an additional worker this period.

Eliminate the expected marginal value of capital and employment from Eqs (12) and (13) using Eqs (9) and (10). This gives expressions for the current marginal value of capital and employment:

$$\begin{aligned} V_k(k(s^t), n(s^t), s^t) &= \frac{\alpha\kappa(s^t)^{\alpha-1} + 1 - \delta}{c(s^t)}, \\ V_n(k(s^t), n(s^t), s^t) &= -\gamma + \frac{(1-\alpha)z(s^t)\kappa(s^t)^{\alpha}}{c(s^t)} \left(1 + \theta(s^t) + \frac{1 - x(s^t) - f(\theta(s^t))}{f'(\theta(s^t))}\right). \end{aligned}$$

Substitute these back into the envelope conditions (12) and (13) to get

$$1 = \beta \sum_{s^{t+1}|s^t} \frac{\Pi(s^{t+1})c(s^t)}{\Pi(s^t)c(s^{t+1})} (\alpha\kappa(s^{t+1})^{\alpha-1} + 1 - \delta) \quad (14)$$

and

$$(1-\alpha)z(s^t)\kappa(s^t)^{\alpha} = \beta f'(\theta(s^t)) \times \left(-\gamma c(s^t) + \sum_{s^{t+1}|s^t} \frac{\Pi(s^{t+1})c(s^t)}{\Pi(s^t)c(s^{t+1})}(1-\alpha)z(s^{t+1})\kappa(s^{t+1})^{\alpha} \times \left(1+\theta(s^{t+1}) + \frac{1-x(s^{t+1})-f(\theta(s^{t+1}))}{f'(\theta(s^{t+1}))}\right)\right). \quad (15)$$

These equations contain the model's main implications. Equation (14) states that the marginal cost of capital, one unit of consumption today, is equal to the expected marginal product of capital next period plus the value of the undepreciated portion of the capital, discounted using the appropriate stochastic discount factor. Equation (15) expresses the tradeoff between recruiting and producing. An additional producer generates $(1-\alpha)z(s^t)\kappa(s^t)^{\alpha}$ units of output in the current period. An additional recruiter yields $f'(\theta(s^t))$ additional workers next period. Each unit of labor supplied reduces utility by $-\gamma c(s^t)$ when measured in units of consumption. In addition, each new recruit permits the planner to put some additional workers into production, each of whom generates $(1-\alpha)z(s^{t+1})\kappa(s^{t+1})^{\alpha}$ units of output. The term on the last line is the number of workers who can be placed into production in period $t+1$ while allowing the firm to maintain its baseline size in period $t+2$. This includes the recruit, some recruiters who can be shifted into production while maintaining the same recruiter-unemployment ratio $t+1$, and a reduction in the recruiter-unemployment ratio at $t+1$ enabled by the continued presence of the new recruit in period $t+2$.

In summary, the solution to the planner's problem is a set of stochastic processes for consumption, capital, the recruiter-unemployment ratio, and employment that satisfies the law of motion for employment in Eq. (7), the resource constraint (8) and the two optimality conditions for investment in physical capital, Eq. (14), and the allocation of labor, Eq. (15).

1.2.3. Decentralization

It is straightforward to decentralize the planner's problem as a search equilibrium with complete markets. A representative household chooses consumption and the purchase of Arrow securities to maximize utility subject to a lifetime budget constraint and a law of motion for employment. A representative firm purchases capital and allocates labor to production and recruiting in order to maximize the present value of profits, discounted using the intertemporal price that clears the asset market. Wages are set by Nash bargaining, where workers' bargaining power is ϕ and the threat point in bargaining is the dissolution of the match. In equilibrium, the goods and asset markets clear and the aggregate recruiter-unemployment ratio is consistent with each firm's labor allocation decision.

Shimer (2010) shows that the conditions that define the equilibrium are nearly unchanged from the social planner's problem. Employment still satisfies the law of motion in Eq. (7) and capital satisfies the resource constraint (8). The first order condition for capital in Eq. (14) also must hold in equilibrium. But the optimality condition for recruiting changes from Eq. (15) to

$$(1-\alpha)z(s^t)\kappa(s^t)^\alpha = \beta\mu(\theta(s^t))\sum_{s^{t+1}|s^t}\frac{\Pi(s^{t+1})c(s^t)}{\Pi(s^t)c(s^{t+1})} \times\left((1-\alpha)z(s^{t+1})\kappa(s^{t+1})^\alpha\left(1+\frac{1-x(s^{t+1})}{\mu(\theta(s^{t+1}))}\right)-w(s^{t+1})\right), \tag{16}$$

where $w(s^{t+1})$ is the equilibrium wage and $\mu(\theta) = f(\theta)/\theta$ is the number of hires per recruiter. Under Nash bargaining, the wage satisfies

$$w(s^t) = \phi(1-\alpha)z(s^t)\kappa(s^t)^\alpha(1+\theta(s^t)) + (1-\phi)\gamma c(s^t). \tag{17}$$

The first term on the right hand side is workers' bargaining power times a measure of the marginal product of labor. This accounts both for the output the worker produces $(1-\alpha)z(s^t)\kappa(s^t)^\alpha$, and for the fact that, if bargaining fails, unemployment increases and the firm must place $\theta(s^t)$ additional workers into recruiting in order to maintain the same recruiter-unemployment ratio. The second term is the firms' bargaining power times the marginal rate of substitution.

One can verify that if the Mortensen (1982)–Hosios (1990) condition holds,[25]

$$1-\phi = \frac{\theta f'(\theta)}{f(\theta)},$$

then Eqs (16) and (17) reduce to the optimality condition for recruiting, Eq. (15). This is possible only if $f(\theta) = \bar{\mu}\theta^{1-\phi}$ and so $\mu(\theta) = \bar{\mu}\theta^{-\phi}$ for some $\bar{\mu} > 0$ and $\phi \in [0, 1]$. Under these conditions, workers' bargaining power is equal to the elasticity of the number of matches with respect to unemployment, and similarly, firms' bargaining power is equal to the elasticity of the matching function with respect to recruiters. This ensures that each firm correctly internalizes the impact of its search on the matching possibility of other firms. Otherwise the equilibrium does not decentralize the planner's problem. In what follows, we refer to wages that decentralize the social planner's solution as "flexible."

[25] Mortensen (1982) was the first to find conditions that guarantee efficiency in a model where matches are generated from individual search decisions. Hosios (1990) derived this condition in the context of a general matching function. See also Diamond (1982b) and Pissarides (1984) for other early contributions to this issue.

1.2.4. Calibration

We cannot solve the model explicitly, and so instead proceed numerically. We calibrate the model using facts about the United States economy, linearize it in a neighborhood of the steady state, and then describe its behavior when hit by shocks. For comparability with much of the existing business cycle literature, we focus on aggregate productivity shocks as the driving force of business cycles, but we also discuss the possible role of other shocks in this framework.

Many of the parameters are standard in the real business cycle literature, but some are specific to the matching model. We start with the more familiar parameters. We think of a time period as one month so as to be able to capture the typical short duration of an unemployment spell. The discount factor is $\beta = 0.996$, just under five percent annually.

We fix $\alpha = 0.33$ to match the capital share of income in the National Income and Product Accounts. We then set $\delta = 0.0028$ per month, which pins down the capital-output ratio at 3.2 along a balanced growth path. This is the average capital-output ratio in the United States since 1948.[26]

We assume productivity has a deterministic trend, $\log z(s^t) = \bar{s}t + s_t$, where $\bar{s}$ is mean productivity growth and s_t follows a first-order autoregressive process,

$$s_{t+1} = \rho s_t + \varsigma \upsilon_{t+1}, \tag{18}$$

where υ_{t+1} is a white noise innovation with mean zero and standard deviation 1. Mean productivity growth is $\bar{s} = 0.0018$, about 2.2 percent per year, consistent with the annual measures of multifactor productivity growth in the private business sector constructed by the Bureau of Labor Statistics.[27] We set the autocorrelation of productivity growth to $\rho = 0.98$ and the standard deviation to $\varsigma = 0.008$. These values are similar to standard calibrations of total factor productivity (Cooley and Prescott, 1995), with an adjustment to account for the fact that time periods are one month long.

We turn next to the parameters that determine flows between employment and unemployment. Shimer (2005b) measures the average exit probability from employment to unemployment in the United States at $x = 0.034$ per month, and we stick with that number here. Initially we assume that it is constant, but we also develop a version of the model with shocks to the employment exit probability.

Although there are many estimates of the matching function f in the literature (see the survey by Petrongolo and Pissarides, 2001), most papers assume that firms create job vacancies in order to attract unemployed workers and so estimate matching functions

[26] More precisely, we use the Bureau of Economic Analysis's Fixed Asset Table 1.1, line 1 to measure the current cost net stock of fixed assets and consumer durable goods. We use National Income and Product Accounts Table 1.1.5, line 1 to measure nominal Gross Domestic Product.

[27] See ftp://ftp.bls.gov/pub/special.requests/opt/mp/prod3.mfptablehis.zip, Table 4. Between 1948 and 2007, productivity grew by 0.818 log points, or approximately 0.014 log points per year. Our model assumes labor-augmenting technical progress, and so we must multiply $\bar{s}$ by $1 - \alpha$ to obtain TFP growth.

using data on unemployment and vacancies. The technology in this paper is slightly different, with firms using workers to recruit workers. Unfortunately we are unaware of any time series showing the number of workers (or hours of work) devoted to recruiting, and so the choice of f is somewhat arbitrary. Still, following much of the search and matching literature, we focus on an isoelastic function, $f(\theta) = \bar{\mu}\theta^{\eta}$, and look at the symmetric case, $\eta = 0.5$. We discuss below the importance of this parameter. To pin down the efficiency parameter in the matching function $\bar{\mu}$, we build on evidence in Hagedorn and Manovskii (2008) and Silva and Toledo (2009). Those papers argue that recruiting a worker uses approximately 4 percent of one worker's quarterly wage, i.e., a recruiter can attract approximately 25 new workers in a quarter, or 8.33 in a month. We use this fact and data on the average unemployment rate to determine $\bar{\mu}$. We proceed in several steps. First, from (7), the steady state employment rate satisfies

$$n = \frac{f(\theta)}{x + f(\theta)}.$$

Setting $n = 0.95$, the average share of the labor force employed during the post-war period, and $x = 0.034$, this implies $f(\theta) = 0.646$ in steady state. Second, the functional form $f(\theta) = \bar{\mu}\theta^{\eta}$ implies

$$\bar{\mu} = \frac{f(\theta)}{\theta^{\eta}} = f(\theta)^{1-\eta}\mu(\theta)^{\eta},$$

where the second equation follows because $\mu(\theta) \equiv f(\theta)/\theta$. From this equation, we set $\bar{\mu} = 2.32$, consistent with $f(\theta) = 0.646$, $\mu(\theta) = 8.33$, and $\eta = 1/2$. Note that this implies that the recruiter-unemployment ratio is $\theta = f(\theta)/\mu(\theta) \approx 0.078$. It follows that the share of recruiters in employment is $\theta(1-n)/n \approx 0.004$, with 99.6 percent of employees devoted to production. Thus in this calibration, the implicit hiring costs are small, at least on average.

Finally, we set the parameter governing the taste for leisure to obtain a five percent unemployment rate along the balanced growth path; this implies $\gamma \approx 0.785$. When we consider alternative calibrations, we vary γ to ensure that the unemployment rate is unchanged.

1.2.5. Results

The search model is useful for developing a notion of unemployment and a theory of worker flows between employment and unemployment. But this section asks whether the model helps to explain other shortcomings of standard business cycle models. What shocks hit the economy? How are they amplified and propagated through time? Why do they create a countercyclical labor wedge?

Shocks We focus in this section on two sources of shocks. One is completely standard in the real business cycle literature (Kydland and Prescott, 1982), the productivity shock

introduced above. The other is special to frictional markets, a shock to the probability of exiting employment. This arguably resembles a "sectoral shift," with many workers losing their job and enduring an unemployment spell before moving elsewhere (Lilien, 1982). Following Blanchard and Diamond (1989), we label the first shock "aggregate" and the second "reallocational," but obviously the names are only suggestive. One important question is whether quantitatively reasonable reallocation shocks are important for the dynamics of employment. We could symmetrically consider a shock to the matching function. In our view, it is implausible to argue that unemployment falls during an expansion because the matching process has exogenously improved, and so we take these shocks off the table.

In principle, one could introduce other shocks to the model. For example, it is straightforward to modify the government budget constraint and resource constraint to introduce stochastic government spending. Trigari (2009) develops a version of a search and matching model with nominal rigidities; in such a framework, monetary policy shocks can also have real effects on output. Of course, one could study both of these shocks in a model without search frictions, and our intuition is that our results comparing models with and without search frictions carry over to these shocks.

Amplification One of our main results is that search frictions dampen the effect of productivity shocks.[28] That is, we compare the volatility of employment and output in a model with search frictions to one without search frictions, i.e., where firms can costlessly adjust employment and wages clear the labor market. We maintain the assumption that leisure is indivisible and so preferences are given by Eq. (6). This is therefore essentially the Hansen (1985) model. The first row in Table 1 considers this frictionless model. The first three columns show the theoretical, infinite sample standard deviation of output, employment, and the consumption-output ratio. The last three columns show a measure more comparable to empirical estimates of these objects. We simulate 402 months of data from our model, compute quarterly averages, take logs and then detrend using an HP filter with smoothing parameter 1600. We show the average results from 1000 such simulations of the model. In both cases output is about 2.2 times as volatile as total factor productivity, while employment and the consumption-output ratio are slightly less volatile than output.[29] Note that in each case the standard deviation of employment and the consumption-output ratio is the same. Indeed, we know from Eq. (5) that, since $\varepsilon = \infty$ and $\sigma = 1$, employment and the consumption-output ratio mirror each other.

[28] Merz (1995) and Andolfatto (1996) assume that recruiting costs are in units of goods rather than labor, which is tantamount to assuming that search costs decrease during periods of high productivity. In that case, search frictions do not substantially dampen the response to productivity shocks. We view a time-intensive model of recruiting as more plausible.

[29] Employment and the consumption-output ratio are less persistent than productivity, while output is about equally persistent. Therefore detrending boosts the volatility of n and c/y relative to productivity but does not much affect the volatility of y relative to productivity.

Table 1 Relative standard deviation of output y, employment n, and the consumption-output ratio c/y in four models.

	Theoretical			Finite sample		
	y	n	c/y	y	n	c/y
Frictionless	2.18	1.39	1.39	2.25	1.85	1.85
Baseline search	1.37	0.15	0.71	1.11	0.18	0.85
Reallocation shocks	1.37	0.18	0.71	1.12	0.27	0.86
Training cost	2.11	1.30	1.33	2.45	1.73	2.07

All variables are expressed relative to the standard deviation of total factor productivity $z(s^t)^{1-\alpha}$. The first three columns show the theoretical, infinite sample standard deviations of monthly variables. The last three columns show detrended quarterly averages based on 402 months of data.

The second row shows our baseline search model. We feed the same shock into the model and affirm that search frictions dampen the response. The standard deviation of detrended output falls by 51 percent, the standard deviation of the detrended consumption-output ratio by 54 percent, and the standard deviation of detrended employment falls most of all, by 90 percent. The theoretical standard deviations fall by a similar magnitude. To the extent that one hoped search frictions would amplify productivity shocks, the results are disappointing. Intuitively, increasing the recruiter-unemployment ratio in response to a positive productivity shock is costly because doing so reduces the effectiveness of each recruiter. This naturally dampens the volatility of employment and hence output. Shimer (2010) verifies that if search frictions are more severe, so each recruiter attracts fewer workers per month, the dampening effect of frictions is even more extreme. Although adding search frictions does not improve our ability to account for the size of employment fluctuations, these results do suggest that it is potentially important to include search frictions when studying employment volatility. To the extent that search frictions significantly dampen the response of employment to shocks, any analysis that abstracts from search may be leaving out a quantitatively important factor.

We next introduce the reallocation shock. We assume $\log x(s^t) = \log \bar{x} + s_{x,t}$, where $\bar{x} = 0.034$ and $s_{x,t}$ follows a linear process,

$$s_{x,t+1} = \rho_x s_{x,t} + \varsigma_x \upsilon_{x,t+1}.$$

We assume the innovation to reallocation, $\upsilon_{x,t+1}$, is white noise with mean 0 and standard deviation 1. We fix $\rho_x = 0.83$ and $\varsigma_x = 0.034$ so as to match the autocorrelation of the employment exit probability and its unconditional standard deviation, as measured in Shimer (2007). For simplicity, we assume that the productivity shock and the employment exit probability shock are uncorrelated, so

$$s_{z,t+1} = \rho_z s_{z,t} + \varsigma_z \upsilon_{z,t+1},$$

where $\upsilon_{z,t+1}$ is independent white noise with mean 0 and standard deviation 1. We

then leave $\rho_z = 0.98$ and $\varsigma_z = 0.008$, as in the model without search frictions.[30] The third row in Table 1 shows that introducing these shocks barely affects the volatility of output or the consumption-output ratio. It raises the theoretical standard deviation of employment by 20 percent and the standard deviation of detrended employment by 50 percent. Nevertheless, employment remains far less volatile than in the frictionless model.

Our calibration strategy tightly pins down most of the important model parameters. The one exception is the elasticity of the matching function, η. Recall that increasing the recruiter-unemployment ratio in response to a positive productivity shock is costly because doing so reduces the effectiveness of each recruiter. The parameter η governs how quickly recruiters' effectiveness falls. When we recalibrate the model with a lower elasticity η, corresponding to the case where unemployed workers are a more important part of the search process, the volatility of each of the three outcomes is significantly dampened. At the extreme case of $\eta = 0$, $f(\theta) = \bar{\mu}$ and so the job finding probability is constant. Absent fluctuations in the fraction of employed workers who become unemployed, x, employment is constant as well, $n = \bar{\mu}/(x + \bar{\mu})$. The search model is equivalent to a real business cycle model with inelastic labor supply.

Conversely, when we raise η, volatility increases. An extreme case is $\eta = 1$, so $f(\theta) = \bar{\mu}\theta$. This implies that a firm must shift $1/\bar{\mu}$ workers away from production in order to hire one new employee. It may be natural to interpret this parameterization of the model as a training cost, rather than a recruiting cost, since the cost depends only on the number of workers hired and not on the availability of unemployed workers. The last row in Table 1 shows that the model with training costs generates almost the same volatility as the frictionless model. Indeed, the volatility of detrended output and the detrended consumption-output ratio actually increase.[31] This result accords with the finding in Mortensen and Nagypál (2007), that a model with training costs generates more volatility than a model with search frictions (see also Pissarides, 2009). Still, the point remains that training costs slightly reduce the volatility of employment compared to a frictionless model, and so in this sense they are a step in the wrong direction.

Finally, it seems worth comparing the predictions of this model with the Lucas and Prescott (1974) search model (see also Alvarez and Veracierto, 1999, 2001). That model focuses on the time-consuming reallocation of workers across labor markets that are continually hit with idiosyncratic productivity shocks. Moving to a new labor market takes one period. This is equivalent to assuming that only unemployed workers are useful in the matching process, $f(\theta) = \bar{\mu}$, and so our analysis here would suggest a dampened

[30] We can also introduce correlation between the shocks, for example by making productivity $s_{z,t+1}$ a function of both the productivity shock $\upsilon_{z,t+1}$ and the reallocation shock $\upsilon_{x,t+1}$. One way to set the correlation between the shocks is to match the correlation between the job finding probability and the employment exit probability, -0.43 in the data. This requires that an increase in reallocation reduce productivity. Doing this does not substantially change our results.

[31] These results appear to be a consequence of time aggregation. When we work directly in a model with a quarterly time period, the standard deviation of all three detrended variables is slightly smaller in the training cost model than in the frictionless model. In any case, the difference in volatility between the two models is small.

response to productivity shocks. Indeed, employment changes in that model only to the extent that the inflow rate into unemployment changes, which we have seen empirically accounts for only a minority of aggregate unemployment fluctuations.[32] Veracierto (2008) attempts to address this by allowing workers to drop out of the labor force, but when he does this, he finds that labor force participation becomes so strongly procyclical in the model that the unemployment rate is weakly procyclical.

Propagation Although search frictions do not amplify shocks, they do affect how they are propagated through time by slowing down the adjustment of employment. To see this, we focus on the first and second order autocorrelation of output growth and employment growth. In United States data, such growth rates are all positively serially correlated: the first and second order autocorrelation of detrended quarterly output growth are 0.23 and 0.08, while the corresponding numbers for detrended quarterly employment growth are 0.19 and 0.17.[33] This recalls the findings of Cogley and Nason (1995), who emphasize the importance of autocorrelations for model evaluation.

Table 2 therefore shows corresponding numbers for the model. The first column shows the theoretical correlation between the growth rate of output from month t to $t + 3$ and its growth rate from $t + 3$ to $t + 6$. The second column shows similar numbers for employment. We see that in the frictionless model, both of these correlations are negative. This reflects the fact that shocks are mean reverting and so output and employment also tend to revert to trend. Adding search frictions boosts this correlation, particularly for employment. On the other hand, reallocation shocks lower the autocorrelation of employment by buffeting it with relatively transitory shocks. Curiously, the training cost model, where matching depends only on recruiters, significantly raises the theoretical autocorrelation of output. These columns suggest that search frictions, particularly the training cost variant, raise the persistence of output and employment.

A direct comparison of the data with the numbers in the first two columns is difficult, both because the data are detrended from a finite sample and because the data are time-aggregated. The third and fourth columns in Table 2 therefore show the first order autocorrelation from similarly time-aggregated model-generated data. Our first result is that, even in the frictionless model, output and employment growth are both positively autocorrelated (0.11). This turns out to be due entirely to time-aggregation. Indeed, even though productivity is mean reverting, the first order autocorrelation of productivity growth is also 0.11. This is because time aggregation raises the first order

[32] Strictly speaking, this is not correct. In the original Lucas and Prescott (1974) model, search is directed, but it takes one period to arrive in the desired labor market. During that period, labor market conditions may worsen substantially, inducing the worker to refuse the job. In practice, this event is rare and so in most cases unemployment lasts for one period.

[33] The output numbers are for real GDP from the National Income and Product Accounts, Table 1.1.6, from 1976Q3 to 2009Q3. The employment numbers are for the fraction of the population who are at work.

Table 2 Autocorrelation of quarterly output and employment growth in four models.

	Theoretical *First order*		**Finite sample** *First order*		**Finite sample** *Second order*	
	y	n	y	n	y	n
Frictionless	−0.03	−0.04	0.11	0.11	−0.11	−0.11
Baseline search	0.03	0.10	0.16	0.22	−0.10	−0.11
Reallocation shocks	0.03	−0.04	0.16	0.10	−0.11	−0.20
Training costs	0.18	0.09	0.27	0.21	−0.13	−0.12

The first two columns show the theoretical, infinite sample autocorrelation of output and employment growth at quarterly frequencies. The next two columns show the first order autocorrelation of detrended quarterly output and employment growth based on 402 months of data. The last two columns show the second order autocorrelation of detrended quarterly output and employment growth based on 402 months of data.

autocorrelation of the growth rate of an autoregressive process; see, for example, Working (1960) for the case of a random walk. This suggests it may be difficult to compare the model with data. Nevertheless, model versus model comparisons are instructive. We again see in the remaining entries in the third and fourth columns that search frictions boost the first order autocorrelation of output and employment growth, particularly in the training cost model, while reallocation shocks moderate the autocorrelation of employment.

Finally, the last two columns in Table 2 look at second order autocorrelations. For a time-aggregated random walk, the first order autocorrelation of growth rates is positive, but the second order autocorrelation is always zero. Therefore we expect that looking at the second order autocorrelation will moderate any issues related to time aggregation. Indeed, we find that the second order autocorrelation of output and employment growth is consistently negative and is basically unaffected by the presence of search costs. None of these search models can therefore generate the persistent positive autocorrelation of output and employment growth that we observe in United States data.

Labor wedge Finally, we turn to the labor wedge. We imagine an economist who understands that the labor supply elasticity is $\varepsilon = \infty$ and the consumption-hours complementarity is $\sigma = 1$. He measures the labor wedge τ using Eq. (5). Table 3 shows the standard deviation of measured $1 - \tau$ and its comovement with output and employment.[34]

In the frictionless model, the labor wedge is always equal to 0, while in the search model it is volatile, and more volatile in the baseline model than the one with training costs. The problem is that $1 - \tau$ is negatively correlated with output and employment in the model, the opposite of the data. This is intuitive. The labor wedge ignores the existence of search frictions, which act as an adjustment cost. Relative to a frictionless model, the adjustment cost dampens fluctuations in employment (Table 1). To rationalize

[34] Note that the choice of the disutility of work γ does not affect the statistical properties of $1 - \tau$, although it determines its average level.

Table 3 Standard deviation of $1-\tau$ and correlation with output and employment in four models.

	Theoretical			Finite sample		
	std. dev.	*correl. with*		*std. dev.*	*correl. with*	
		y	n		y	n
Frictionless	0	—	—	0	—	—
Baseline search	0.56	−0.73	−0.96	0.68	−0.98	−0.93
Reallocation shocks	0.38	−0.72	−0.77	0.69	−0.96	−0.55
Training costs	0.06	−0.46	−0.40	0.53	−0.61	−0.47

The first three columns show the theoretical, infinite sample standard deviation of $1-\tau$ and correlation with output and employment. The last three columns show the standard deviation of the detrended quarterly labor wedge and its correlation with detrended output and detrended employment based on 402 months of data.

this in a frictionless model, we need to assume that the tax on labor rises during every expansion. But this is exactly the opposite of what we observe in the data. Again, the negative correlation between the labor wedge and output or employment is smaller in the training cost version of the model, but the counterfactual implication remains highly significant.

Other moments One can examine other moments in the model, for example the behavior of consumption and investment. Doing so reaffirms earlier work emphasizing that the presence of search frictions does not substantially modify the behavior of a business cycle model (Merz, 1995; Andolfatto, 1996). The main effect of search frictions is instead to dampen the response of the economy, and especially the labor market, to aggregate shocks. While this is disappointing, it is worth stressing that the model has some advantages over the baseline business cycle model, namely the introduction of unemployment and aggregate worker flows.

1.3. Rigid wages

The concern that search models do not generate substantial fluctuations in unemployment was first voiced in Shimer (2005b), albeit in a model with linear utility and no capital; see also Costain and Reiter (2008). The body of Shimmer's paper focused on a model with Nash bargaining that satisfied the Mortensen-Hosios condition, i.e., what we have defined as a flexible wage model with outcomes equivalent to the social planner's problem. While wages that decentralize the social planner's solution may be a useful benchmark, the assumption is not obviously more plausible than a myriad of possible alternatives.[35] In his conclusion, Shimmer argued that wage rigidities—wages that are

[35] A number of papers assume that firms post wage offers and workers can direct their search towards their preferred offer, as in the competitive search literature (Montgomery, 1991; Peters, 1991; Moen, 1997; Shimer, 1996; Acemoglu and Shimer, 1999; Burdett et al., 2001; Mortensen and Wright, 2002). The equilibrium of that model coincides with the social planner's solution, and so in our terminology wages are flexible. Models of wage rigidities therefore typically assume either that firms cannot commit to wages or that workers cannot direct their search.

less procyclical than those which decentralize the social planner's solution—may help to resolve the "unemployment volatility puzzle."[36] This section starts by reviewing the subsequent theoretical literature on wage rigidities, then discusses papers that attempt to measure whether wages are rigid in reality, and concludes by considering whether one needs a model with search frictions to analyze wage rigidities.

1.3.1. Theory

Hall (2005) was the first paper to quantify the possibility of wage rigidities creating volatile unemployment in a search model. He replaced the Nash bargaining assumption, analogous to Eq. (17) here, with a restriction that wages do not move in response to aggregate productivity shocks. A temporary increase in productivity therefore raises the revenue from hiring workers without raising the cost. This induces firms to recruit more workers, which in turn lowers the unemployment rate. Hall (2005) established that this is indeed a powerful amplification mechanism.

An important insight of Hall (2005) was that the wage negotiation between a matched worker and firm is a zero-sum game. The marginal rate of substitution between consumption and leisure is strictly less than the marginal product of labor in a search equilibrium, with the difference representing a match-specific rent due to the existence of the search friction. It follows that there is a range of wages that a worker is willing to accept and a firm is willing to pay. This has two implications. First, search models with rigid wages do not suffer from the Barro (1977) critique of unemployment in the implicit contracts literature (e.g. Baily, 1974; Gordon, 1974; Azariadis, 1975), that inefficient layoffs arise only because matched workers and firms fail to exploit some of the bilateral gains from trade.[37] Second, wage rigidities in existing employment relationships are inconsequential, so long as they do not lead to inefficient separations (Shimer, 2004).[38] Instead, the recent search literature has focused on how wage rigidities affect firms' incentive to create job vacancies and to recruit new employees, leading to fluctuations in the job finding probability. Given the empirical evidence that the job finding probability declines sharply and remains low long after the initial recessionary shock, this emphasis seems reasonable.

[36] This recognition of the central role of wage determination in search models is not new. In his review of the first edition of Pissarides (2000), Mortensen (1992, p. 166) noted that "the fact that alternative rules of wage determination may have different implications is an important neglected topic .... Unlike the Walrasian theory, there is no unique concept of equilibrium price inherent in the theory of markets with transactions costs. Wages must be determined by some form of bargaining and the implications of the model are generally sensitive to which bargaining solution is imposed." Caballero and Hammour (1996) also stress that how match-specific rents are appropriated may be important for business cycle fluctuations.

[37] With two-sided asymmetric information, inefficient separations may be a necessary feature of equilibrium. See Ramey and Watson (1997) for an example where layoffs arise in a search model because of endogenous limits on contracting.

[38] With incomplete markets, wage rigidities in existing employment relationships matter because they affect the value of the relationship and so affect job creation; however, Rudanko (2009) finds that this effect is quantitatively small.

The subsequent theoretical literature on wage rigidities in search models can largely be divided along two dimensions. First, some papers attempt to provide a deep foundation for the rigidity, while others pursue a more ad hoc approach. Second, in some papers the wage rigidity is intrinsically static, while in others it introduces an additional state variable.[39]

The simplest wage rigidity models are ad hoc and static. One example is Hall (2005), who assumed wages are fixed forever. Hall presents a simple bargaining game in which any wage between the marginal rate of substitution and the marginal product of labor is an equilibrium. He then argues that "a constant wage rule may be interpreted as a wage norm or social consensus." (p. 56) Blanchard and Galí (2008) also impose an ad hoc, static wage rule. Generalizing Hall (2005), they assume that the wage is proportional to productivity, but the constant of proportionality is smaller than 1. Thus when productivity is high, the gap between the marginal product of labor and the wage is large, again encouraging firms to hire. The free proportionality parameter affects the extent of wage rigidity and hence the volatility of unemployment. It is worth stressing that even with this free parameter, the model is testable. For example, Hall (2009) examines the implications of a rigid wage model for the behavior of hours per worker, under the assumption that workers and firms negotiate hours efficiently, even if total compensation is rigid. He finds the model can eliminate the cyclicality of the labor wedge.

Other authors present more sophisticated arguments for why wages are rigid. An early example is Hagedorn and Manovskii (2008), although it is worth noting that the authors do not interpret their paper as one with wage rigidities. They calibrate the Nash bargaining parameter using information that wages move less than one-for-one with productivity, which gives them a small value for the workers' bargaining power ϕ in (17). This significantly amplifies productivity shocks relative to the baseline search model. To understand why, recall that if $\phi = 1 - \eta$, so that the Mortensen (1982)–Hosios (1990) condition is satisfied, the equilibrium is equivalent to the solution to the social planner's problem. Recall that we also found that when $\eta = 1$, i.e., the training cost model, search frictions do not much dampen productivity shocks. It turns out that when $\phi = 0$ but $\eta < 1$, the results are similar although not quite as strong; search frictions still dampen productivity shocks, but not as much as in the baseline search model. For example, if $\phi = 0$ and $\eta = 0.5$, the standard deviation of detrended employment is 1.3 times the size of the productivity shock, far more than in the baseline search model but somewhat less than the value of 1.7 in the training cost model. The autocorrelation of output and employment growth are also essentially the same as in the training cost model and the procyclical labor wedge is unaffected as well.

Another approach to static wage rigidities comes from reexamining the threat point when bargaining. In Eq. (17), we assumed that a breakdown in bargaining led to the

[39] There may also be institutional reasons, such as unions, why wages are rigid. We do not know of any recent attempts to understand whether wage rigidities at business cycle frequencies are consistent with unions' objective function.

dissolution of a match. Hall and Milgrom (2008) argue that a worker and firm are likely to continue bargaining even if agreement is not immediate. Therefore the threat point when bargaining is delay, not breakdown. This small change has a big effect on the equilibrium, since the value of delay—say the worker's time and the firm's foregone production—is less cyclical than the value of dissolution. Such a model can potentially generate more volatility than the frictionless benchmark, and a countercyclical labor wedge.

A third approach to static wage rigidities is based on asymmetric information. Kennan (2010) explores what happens if workers are unable to observe the productivity of their match. He shows that under some conditions the information rent accruing to firms is procyclical, effectively generating rigid wages and amplifying the impact of productivity shocks. Other information frictions, such as the need to pay a high wage that keeps workers from shirking (Shapiro and Stiglitz, 1984; MacLeod and Malcomson, 1989) may also be important for wage setting; see Costain and Jansen (2009) for recent work integrating efficiency wages into a search model.

These static stories amplify shocks but typically do not propagate them. Models where wages are backward looking can do both. Again, the simplest models are ad hoc. In Blanchard and Galí (2007), the current wage is a weighted average of the previous period's wage and the marginal rate of substitution between consumption and leisure. In Shimer (2010, Chapter 4), it is a weighted average of the previous wage and the current wage that would prevail if there were Nash bargaining, which significantly propagates shocks without much affecting the comovement of wages and labor productivity.

Gertler and Trigari (2009) provide a deeper theory for why wages may be backward looking. They assume that workers and firms only periodically negotiate, bargaining so as to satisfy the Nash solution and fixing the wage until the next opportunity to renegotiate. Crucially the negotiated wage applies not only to the firm's existing workers, but also to any new workers it might hire. Thus firms that last negotiated their wage prior to an adverse productivity shock will have little incentive to recruit new workers following the shock. They again show that this amplifies the effect of shocks on the labor market with little consequence for other macroeconomic outcomes.

1.3.2. Evidence

This theoretical literature points to an obvious empirical question: are wages in reality rigid? Some recent papers have in fact argued that wages in new matches are flexible, as evidenced by the fact that they are as volatile as labor productivity (Pissarides, 2009; Haefke et al., 2008); however, this evidence is also consistent with a rigid wage model. To understand why, it is useful to step back and think about an otherwise frictionless model where, for some reason, wages are above the market-clearing level. In this case, firms set the level of employment so that the marginal product of labor is equal to the wage. With a Cobb-Douglas production function, it follows that labor productivity,

i.e., the average product of labor, is proportional to the wage; see Eq. (4).[40] This is true regardless of the source of shocks, and indeed regardless of whether wages are rigid. That is, in the absence of search frictions, the observation that wages are as volatile as labor productivity is uninformative about whether wages are rigid. This argument does not exactly carry over to a model with search frictions, but quantitatively it is not far off. Shimer (2010, Chapter 4) reports that the labor share, $w_t h_t / y_t$, is nearly constant in a model where wages are extremely backward looking. Devising appropriate tests for whether wages are rigid remains an important issue for future research.

1.3.3. Why search?

A final issue is whether search models are the right framework for thinking about wage rigidities. It is clear that search frictions provide one possible explanation for why wages are rigid: there is a gap between the wage that workers will accept and firms will pay, and any wage between those bounds is in some sense reasonable. But anything that creates match-specific rents, such as training costs, match-specific capital, labor unions, and collusion among employers, also creates rents.[41]

A natural question is therefore whether there is any reason to prefer search over other models of rents. An important feature of search models is that it is impossible for a worker and a firm to contract on the division of rents before the rents are created, since they have not yet met. In contrast, it is in principle possible to write contracts that divide the ex post rents from training costs, match-specific capital accumulation, or private information. Of course, there may be limits to that contracting. For example, limits on workers' ability to commit to stay in an employment relationship may mean that a firm will only invest in a worker's human capital if the worker can post a bond; and borrowing constraints may prevent bond-posting. Alternatively, employees' morale may constrain the firm and prevent it from cutting wages during downturns (Bewley, 2002). Still, search offers a potentially important explanation for why match-specific rents are not divided efficiently.

Finally, one might think about wage rigidities in a model without any match-specific rents. At a crude level, one could impose a wage above the market-clearing level in an otherwise competitive framework. The aggregate implications for employment, unemployment, and the labor wedge would be very similar to what comes out of a search model. But this approach seems unsatisfactory to us, since the wage rigidity would necessarily have to be ad hoc, and hence potentially not robust to different policy interventions. To develop a deep theory of wage rigidities, one needs a model that sheds light on the forces that prevent wages from adjusting to clear the labor market. The trading frictions inherent in search models seem a promising way of understanding these

[40] For evidence in support of this theory, see Galí et al. (2007), who attribute all the volatility in the labor wedge to the "wage markup," the ratio of the real wage to the marginal rate of substitution in Eq. (3). They find almost no volatility in the "price markup," the ratio of the marginal product of labor to the real wage in Eq. (4).

[41] For a discussion of these sources of rents, see the chapter by Manning in this volume.

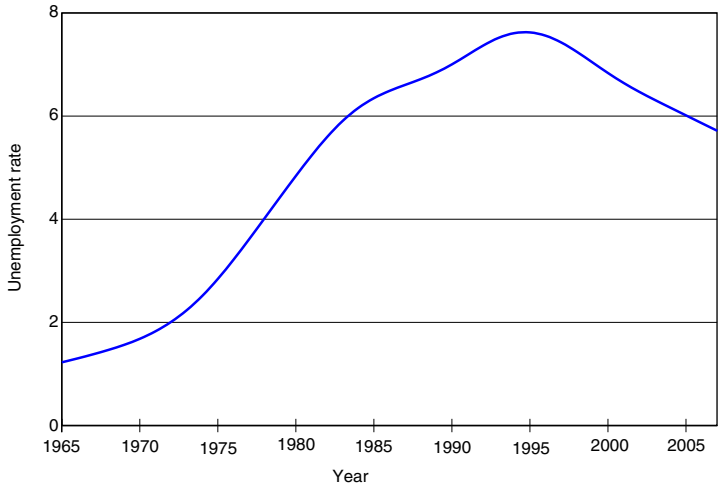


Figure 14 ***The line shows the trend unemployment rate in the OECD.***

forces. To the extent that substantial heterogeneity across workers and jobs makes search frictions more pronounced in the labor market than in most other markets, this may help us to understand why rigidities are more important in the labor market as well.

2. TRENDS

We now shift our attention from cyclical fluctuations to long-run trends. The persistent, widespread, but unequal increase in unemployment across OECD countries in the 1970s and 1980s motivated a substantial body of research that sought to understand why different countries experienced different outcomes. This section reviews some of the key features of the low-frequency data and then examines how search theory has been used to understand these trend changes in labor market outcomes.

2.1. Facts

2.1.1. Unemployment rate

We start by presenting two key features of trend changes in unemployment across the OECD since 1965.[42] First, almost all countries experienced a single peak in trend unemployment during this period. Second, there was substantial heterogeneity in the extent of the increase in trend unemployment and the timing of the peak.

Figure 14 displays the average trend unemployment rate for 17 OECD countries from 1965 until 2007, where the trend is defined as an HP filter with smoothing parameter

[42] Layard et al. (2005) and Blanchard (2006) are recent contributions that also summarize some facts about unemployment evolutions in the OECD.

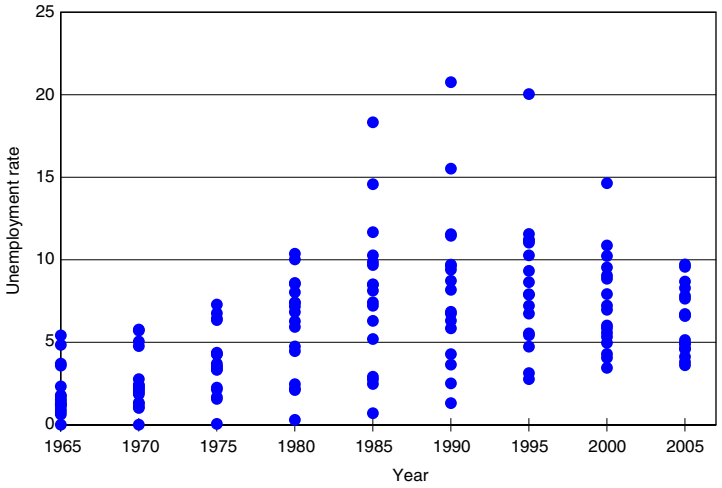


Figure 15 ***Each dot shows the trend unemployment rate in one OECD country at five year intervals.***

100.[43] As the picture shows, trend unemployment increased steadily from the beginning of the sample until the mid 1990s, ultimately rising by a factor of six. This increase was far larger than the changes associated with business cycle fluctuations.

The subsequent decline in trend unemployment was equally important, although even in 2007 the level is more than three times higher than it was in the mid 1960s. Researchers studying unemployment through the mid 1990s took as a starting point that one needed to understand the factors that caused a permanent increase in unemployment. But Fig. 14 suggests that the key fact to explain is not a permanent increase in unemployment but rather a long-lived temporary increase.

A particularly striking feature of the data is the heterogeneity in unemployment evolutions across countries. Figure 15 plots the distribution of unemployment rates at five year intervals starting in 1965. An interesting feature of this figure is that as mean unemployment increased from 1965 to 1995, there was a marked increase in the dispersion of unemployment rates as well. The subsequent decrease in unemployment was accompanied by a corresponding decrease in dispersion. In fact, as shown in Fig. 16, the standard deviation of unemployment rates in 2007 is roughly the same as it was in 1965 despite the fact that mean unemployment is higher.

Next we look at trend unemployment in each of the individual countries. We note upfront that these figures show two striking features. First, in almost every country the evolution of trend unemployment followed a single peaked shape, similar to that found for the cross-country average. Second, although the qualitative shape of trend

[43] Throughout this section, all OECD averages are unweighted by population. For further details on the data source and countries, see footnote 7.

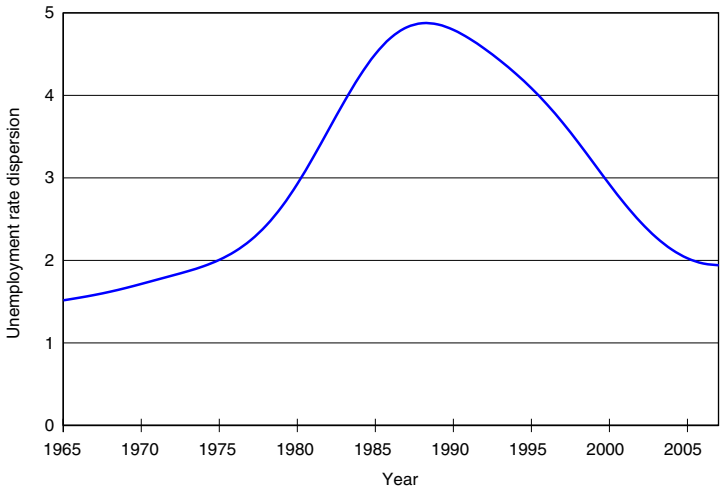


Figure 16 ***The line shows the standard deviation of the trend unemployment rate across OECD countries.***

unemployment is the same for all countries, the extent of the increase from initial level to peak level and the timing of the peak varied significantly across countries.

We begin by displaying in Fig. 17 the evolution of trend unemployment in France, Germany, the United Kingdom and the United States. This figure illustrates these two general features. While each country followed the same single peaked shape, the magnitude of the increase from initial level to peak level was four percent in the United States but more than nine percent in the other countries. And while the United States reached its peak in the early 1980s and then rose again modestly at the end of the sample, the United Kingdom peaked in the mid 1980s, followed by France in the mid 1990s, and Germany near the end of the sample. Note that there is no strong relationship between the timing of the peak and the extent of the initial increase. For example, France and the United Kingdom had roughly the same increase up to the peak, but the United Kingdom reaches its peak roughly ten years earlier.

Some of our subsequent analysis will focus on these four economies due to their size and importance in various debates about unemployment. But first it is interesting to note that the features observed for these four countries also hold for every other country in our sample except Portugal. Figures 18–20 show the evolution of trend unemployment for thirteen OECD economies.

To quickly summarize, while in all of the figures we basically see dynamics that follow a single peaked profile, the timing of the peak and the extent of the increase prior to the peak differ quite dramatically across countries. Similarly, the change in unemployment from the beginning to the end of the period was very different across countries. Given that the dynamics of unemployment appear to be persistent but not permanent, it is

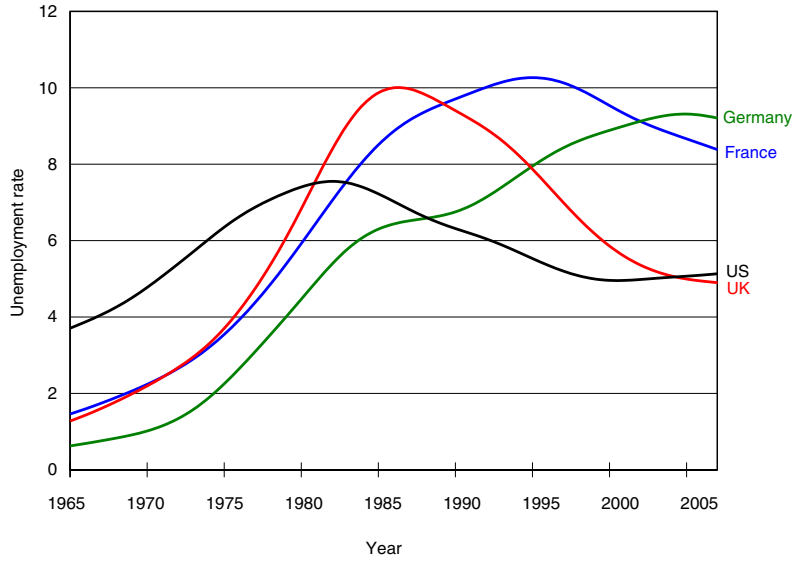


Figure 17 *The four lines show trend unemployment in Germany, France, the United States, and the United Kingdom.*

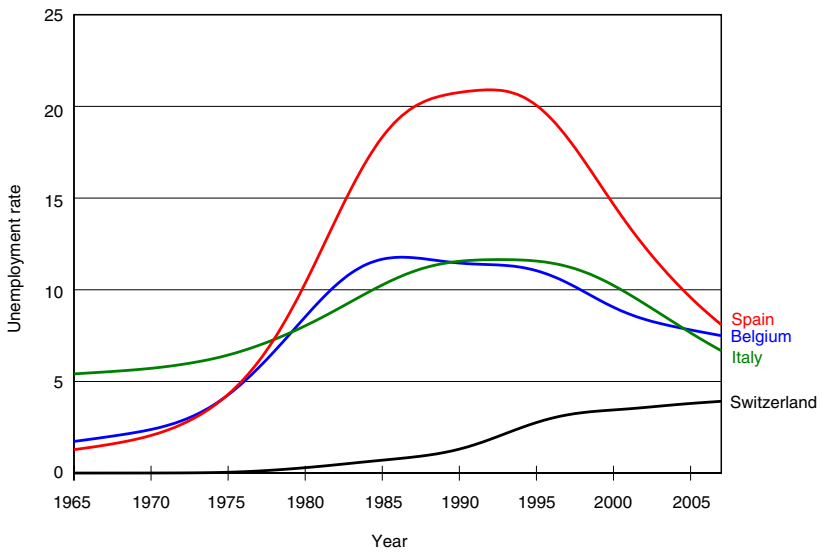


Figure 18 *The four lines show trend unemployment in Spain, Belgium, Italy, and Switzerland.*

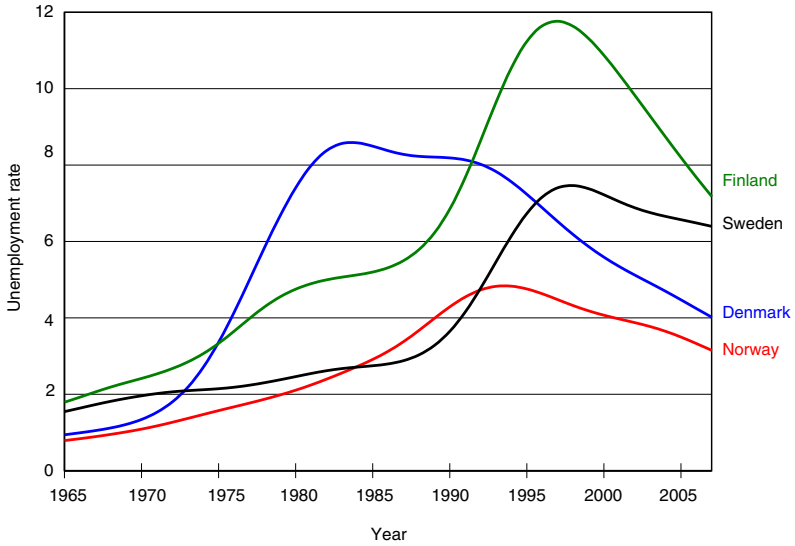


Figure 19 ***The four lines show trend unemployment in Finland, Sweden, Denmark, and Norway.***

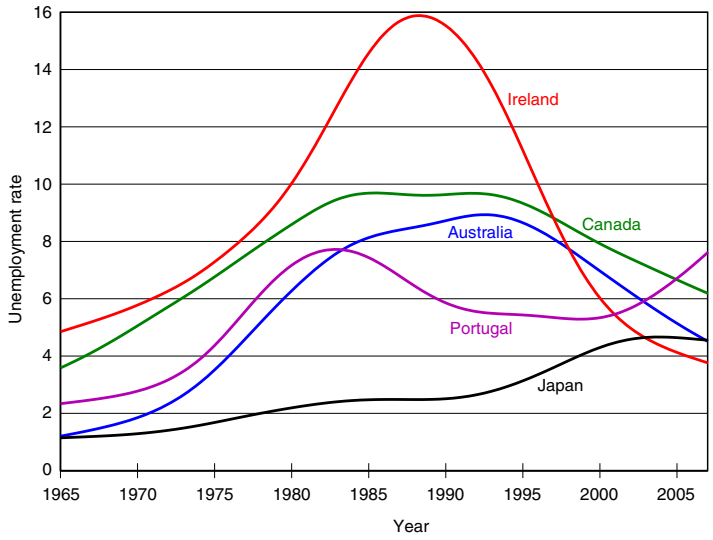


Figure 20 ***The five lines show trend unemployment in Ireland, Canada, Australia, Portugal, and Japan.***

difficult to say at this point whether these differences will continue to change in the future. For example, in countries such as the United States and Ireland, unemployment is basically the same at the beginning and the end of the period. In contrast, in countries such as France and Germany, unemployment is substantially higher at the end of the

period. But given that France and Germany reached their peak unemployment levels much later than the United States, it is unclear whether trend unemployment in these countries will continue to decline.

2.1.2. Total hours

The divergent evolution of the unemployment rate across countries invites an analysis of the sources of these differences. But before we discuss this, we place these trend unemployment changes in a broader context. In our discussion of United States cyclical fluctuations, we concluded that changes in hours per worker account for about a third of the overall volatility in total hours and that the labor force participation rate is comparatively acyclic. This means that changes in unemployment over the business cycle capture a great deal of the change in total hours worked over the business cycle, and so to a first approximation, understanding cyclical fluctuations in total hours amounts to understanding the movement of workers between employment and unemployment. To the extent that search theory informs us how workers move between unemployment and employment, search theory could potentially play a key role in understanding movements in aggregate labor market outcomes at business cycle frequencies.

We have already established that there are large low frequency changes in unemployment over time, and so now we examine whether these are the dominant source of low frequency changes in aggregate labor market outcomes. We show that low-frequency changes in unemployment account for a relatively small share of the movement in total hours for the period from 1965 to 1995. They have become relatively more important since 1995, because the magnitude of movements in total hours has diminished. Additionally, cross-sectional differences in unemployment currently account for a very small fraction of the overall dispersion in total hours. This suggests that, although search theory has the potential to shed light on the forces that shape the large low frequency movement in unemployment over time, it is not likely to be of first order importance in understanding changes in total hours.

We start by looking at the evolution of the cross-country average of trend total hours for the 17 countries in our sample. Figure 21 shows a dramatic decline in total hours between 1965 and 1995. The magnitude of the drop exceeds 15 percent, which is again much larger than the decline in total hours associated with business cycle fluctuations. There is a striking similarity between the time series data for average unemployment and average total hours. Whereas unemployment increased until the mid 1990s and declined thereafter, total hours decreased until the mid 1990s and increased thereafter. And while unemployment displays a net increase over the entire period, total hours displays a net decrease. Based on a cursory look at the patterns, one might be tempted to conclude that the changes in total hours and changes in unemployment are just two ways of describing the same phenomenon. However, a somewhat closer look reveals

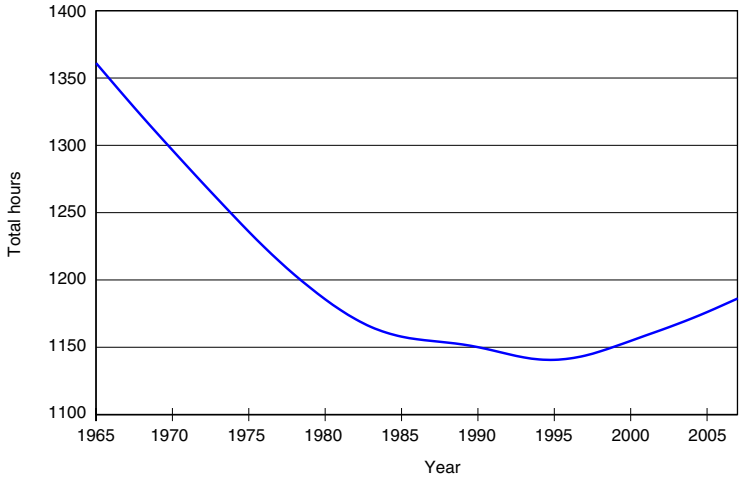


Figure 21 ***The line shows the OECD average of country trends in total hours.***

some important differences. For example, whereas the decrease in total hours was almost complete as of 1985, the unemployment rate continued to increase sharply after 1985.

A simple decomposition allows a more quantitative assessment of the importance of unemployment rate changes. Our measure of total hours (H) is the product of hours per worker (h) and the employment-population ratio (E/P). And the employment-population ratio can itself be expressed as the product of the participation rate (PR) and one minus the unemployment rate ($1 - UR$). To see this note that if we let the stock of employed, unemployed and total population be denoted by E, U, and P respectively, then $PR = (E + U)/P$ and $1 - UR = E/(E + U)$. Total hours can then be expressed as the product of three terms:

$$H = h \cdot PR \cdot (1 - UR).$$

The contribution of changes in unemployment to changes in total hours is accounted for by the third term. Figure 22 plots the time series for the trend components of the cross-country averages for each of these three series, with each value expressed relative to its 1965 value in order to facilitate comparisons. By examining the relative change in each of these three terms over time we can assess the relative importance of each component.

The figure shows that over the entire period, the increase in the unemployment rate reduced total hours by about four percent. At its peak, in the mid 1990s, the contribution was around seven percent. While this is large relative to changes at business cycle frequencies, it is much smaller than any of the other trend changes that took place in the labor market over this same time period. For example, the increase in labor force

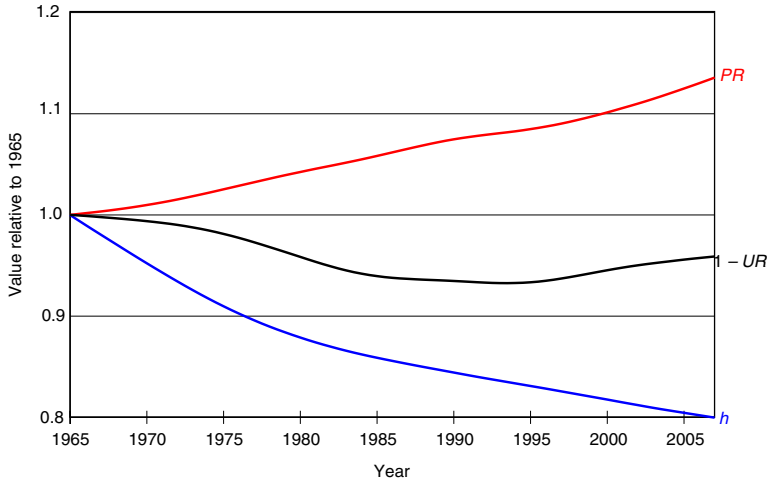


Figure 22 ***The three lines show the labor force participation rate*** PR***, the employment-labor force ratio*** $1 - UR$***, and hours per worker*** h***. Each is the OECD average of country trends.***

participation raised total hours by almost 15 percent, while the decrease in hours per employed worker lowered total hours by 20 percent.

The large trend changes in participation and hours per worker are probably not associated with search frictions. The increase in participation is due entirely to the increased participation of women, and it seems unlikely that search is a key factor in understanding the widespread increase in female participation.[44] And the decrease in hours per worker is accounted for by increases in vacation days and statutory holidays, decreases in the length of the full-time work week, and increases in part-time work. Once again, search frictions do not seem to be a key element in explaining these trends. We conclude that search is probably not a key element of the explanation for the dramatic decline in hours worked over the entire period since 1965.[45]

To pursue this a bit further, we focus on four individual countries—France, Germany, the United Kingdom and the United States. We begin in Fig. 23 by displaying the series for total hours. All three European countries experienced a very significant drop in total hours over this period, ranging between 25 and 35 percent. In contrast, the change in total hours for the United States between 1965 and 2007 was relatively modest.

There are again some qualitative similarities between the evolution of total hours and unemployment rates for these four economies. The United States had relatively little change in both its unemployment rate and total hours from 1965 to 2007, though in

[44] For example, prominent papers in this literature include Galor and Weil (1996), Goldin and Katz (2002), Jones et al. (2003), Greenwood et al. (2005), Olivetti (2006), and Attanasio et al. (2008), and none even mentions search as an important element to consider.

[45] See Pissarides (2007) for an analysis that jointly considers evolutions in unemployment and total hours.

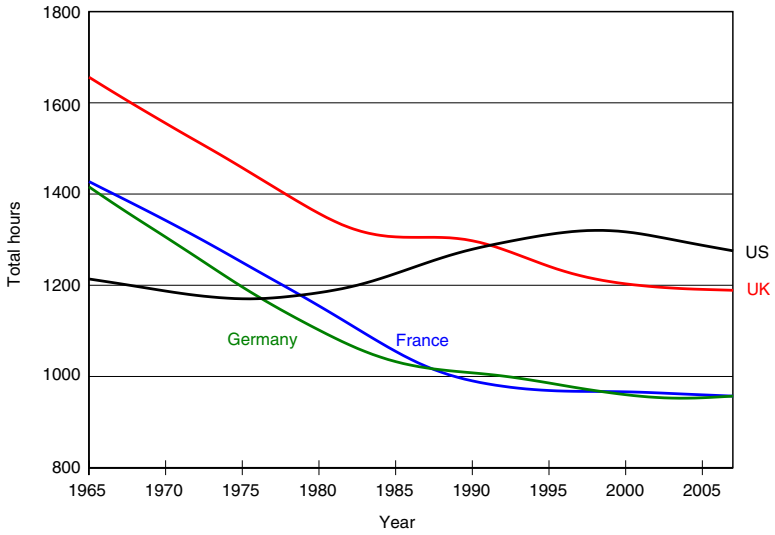


Figure 23 ***The four lines show trend total hours in Germany, France, the United States, and the United Kingdom.***

each case there are some low frequency movements between the two endpoints. And for each of the three other countries, there was a net increase in the unemployment rate and a net decrease in total hours over the period.

Once again, these qualitative comparisons at the individual country level might lead one to suspect that changes in total hours were dominated by changes in unemployment. Figures 24–27 decompose low frequency movements in total hours for each of these four countries. While the exact numbers vary a little across countries, these figures confirm the earlier conclusion reached on the basis of cross-country averages—the participation and hours per worker margins were collectively much more important than the unemployment margin in accounting for changes in total hours. In particular, whereas total hours fell between 25 and 35 percent for the three European economies, the decrease accounted for by changes in the unemployment rate was only between four and nine percent.

A related but distinct calculation is to ask how important cross-sectional differences in unemployment are in accounting for cross-sectional differences in total hours. More precisely, we ask what would total hours be in a country relative to the United States if we were to move individuals between employment and unemployment so as to give all countries the same unemployment rate as the United States, but leave hours per worker and labor force participation unchanged. Figure 28 reports the results of such an exercise based on the 2005 cross-section. Consistent with the earlier calculations, we see that differences in unemployment account for differences in total hours on the order of five percent or less. While differences of this magnitude are quantitatively important from a

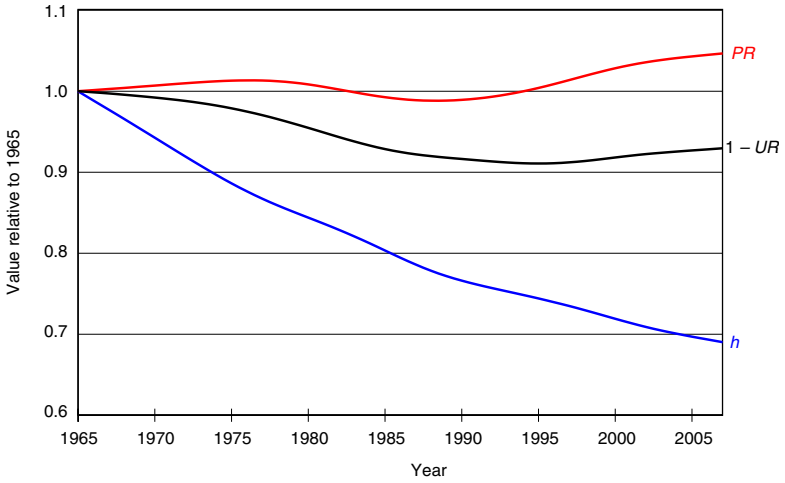


Figure 24 ***The three lines show the labor force participation rate PR, the employment-labor force ratio $1 - UR$, and hours per worker h in France.***

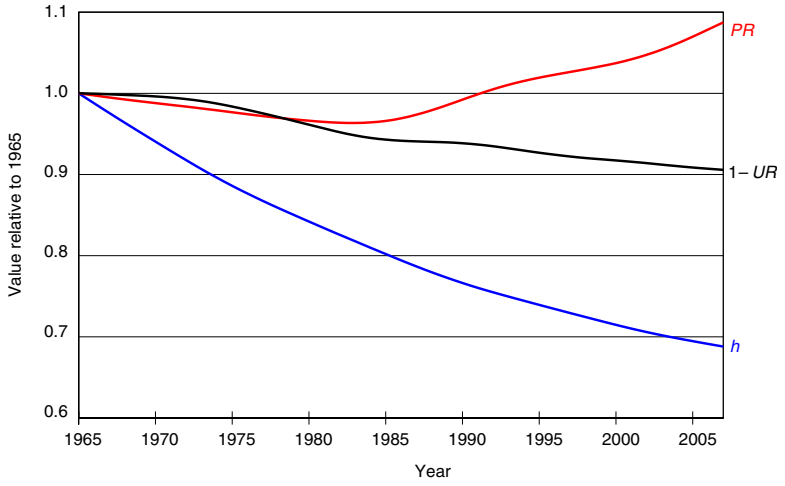


Figure 25 ***The three lines show the labor force participation rate PR, the employment-labor force ratio $1 - UR$, and hours per worker h in Germany.***

business cycle perspective, they are relatively small in the context of understanding the cross-sectional dispersion in total hours across countries.

2.1.3. Unemployment inflows and outflows

One key feature of search models is that they make predictions about flows into and out of unemployment. In our earlier analysis of business cycle fluctuations, we argued that

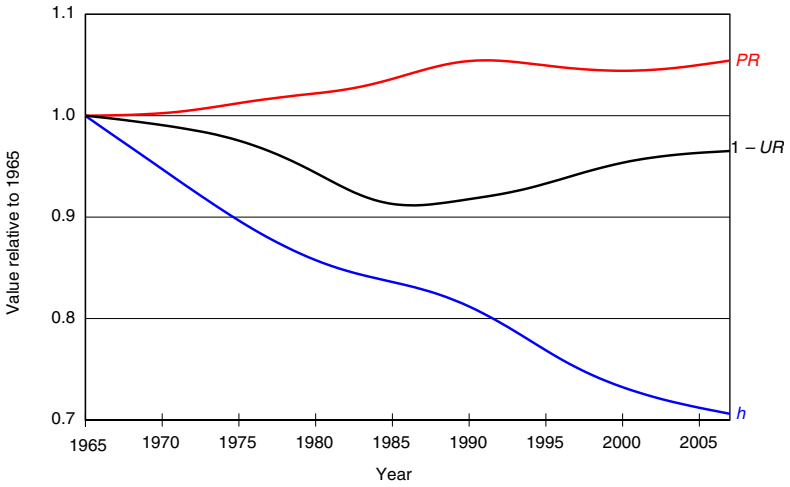


Figure 26 ***The three lines show the labor force participation rate PR, the employment-labor force ratio $1 - UR$, and hours per worker h in the United Kingdom.***

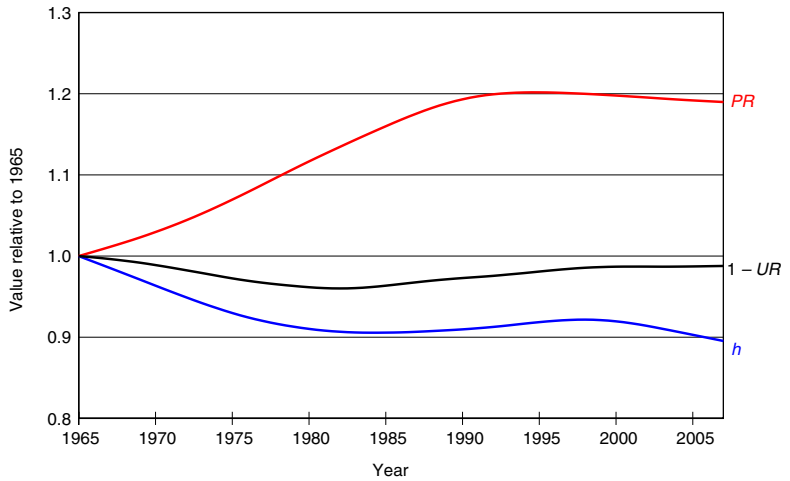


Figure 27 ***The three lines show the labor force participation rate PR, the employment-labor force ratio $1 - UR$, and hours per worker h in the United States.***

recessions are characterized by a short, sharp spike in the inflow rate into unemployment and a persistent decline in the outflow rate. We are interested in knowing whether persistent changes in the unemployment outflow rate also accounted for the substantial increase in the unemployment rate from 1965 to 1995 and the subsequent reversal. Unfortunately, data availability limits the extent to which one can readily carry out such

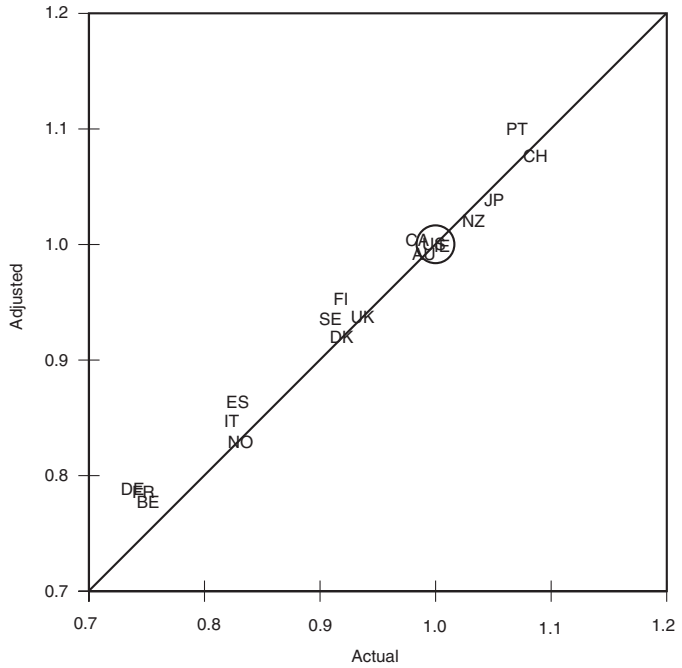


Figure 28 ***Total hours relative to the United States 2005, actual and adjusted for unemployment.***

an analysis for a large set of countries over a long time period. Nonetheless, there are data that can shed some light on this issue, and recent work has made some headway in producing estimates for several countries. In this section we summarize this evidence.

The evidence that we present here supports the following three conclusions. First, there are large differences in unemployment inflow and outflow probabilities across countries that are not related to differences in unemployment rates. Second, in terms of accounting for low frequency changes in unemployment, changes in both inflow and outflow probabilities played a significant role. Third, there does not appear to be a systematic pattern regarding the importance of changes in inflows and outflows that holds across countries. Moreover, for some countries the relative importance of these two flows changes over time.

Our main source of worker flow data is the OECD, which publishes the distribution of unemployment duration for the current stock of unemployed workers. The coverage is incomplete, starting at 1976 in some countries but not until 1983 for many other countries. For the most part, therefore, these results apply to the post 1980 period. We emphasize that many factors can influence unemployment duration distributions, including the prevalence of switches between unemployment and inactivity, the role of temporary layoffs and temporary jobs, the demographic and industrial composition of the workforce,

etc. The development of high quality, comparable time series measures of worker flows which control for variations in these factors is an important issue for future work.

We use data on employment, unemployment, and the fraction of workers who have been unemployed for less than one month to make some inferences about the inflow and outflow probabilities over time for a cross-section of OECD countries. A more thorough analysis can be found in Elsby et al. (2008), and we refer the reader to that paper for a detailed discussion of some key issues. To construct the unemployment outflow (or job finding) probability, we use a version of Eq. (1), but impose steady state, i.e., $u_t = u_{t+1}$ and $u_t^{<1} = u_{t+1}^{<1}$. Thus the unemployment outflow probability is just $F_t = u_t^{<1}/u_t$, the fraction of unemployed workers with duration less than one month. Intuitively, in steady state the outflow of new workers balances the inflow, and so the number of workers unemployed for less than a month is equal to the unemployment outflow. We use a similar approach to construct the unemployment inflow probability. After correcting for time-aggregation, this gives $X_t = 1 - (1 - F_t)^{u_t/e_t}$.[46] Elsby et al. (2008) show how one can relax the steady state assumptions by using additional information on the stock of unemployed by duration. In the United States case, imposing steady state scarcely affects the behavior of these two time series. More generally, for the points we emphasize here, the steady state assumption is relatively innocuous, and so we present simple estimates that use that assumption.

We focus on data for four countries that illustrate our key points, France, Sweden, the United Kingdom and the United States. Figure 29 presents a scatter plot of the trend unemployment rate against the trend unemployment outflow probability for these four countries. Two features are apparent. First, with the exception of the United States, there is negative correlation between trend changes in the unemployment rate and trend changes in the unemployment outflow probability. We will consider this in more detail later on in this section. Second, even at a given level of trend unemployment, there are dramatic differences in the unemployment outflow probability across countries. For example, all countries in this group experienced a trend unemployment rate of six percent at some point during the sample period. But at this level of trend unemployment, the trend unemployment outflow probability varied from around five percent in France to around forty percent in the United States. Even if we focus on the three European countries, the differences were still enormous, with the probability for the United Kingdom three times as high as the probability in France, and Sweden 1.5 times as large again.

Figure 30 compares the trend unemployment rate with the trend unemployment inflow probability for the same group of countries. Once again, two features are apparent. First, for all countries except the United Kingdom, there is a positive relationship between changes in the trend unemployment rate and changes in the trend unemployment inflow probability. Second, even at a given unemployment rate, there were very

[46] For a derivation of this formula, see Shimer (2007).

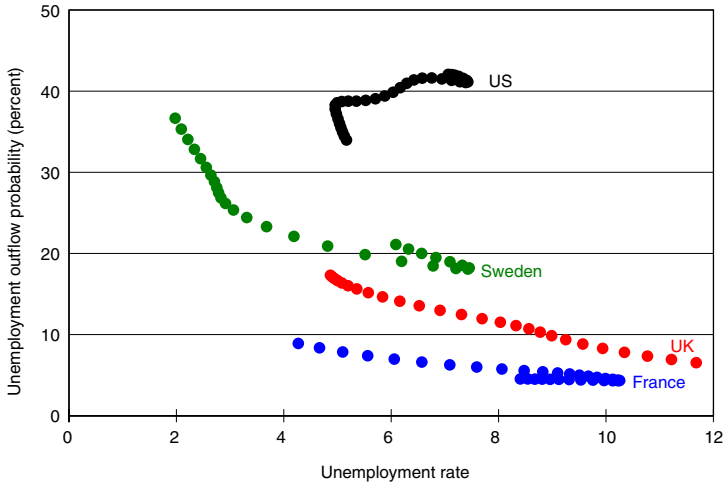


Figure 29 ***Unemployment rate and unemployment outflow probability in France, Sweden, the United Kingdom, and the United States.***

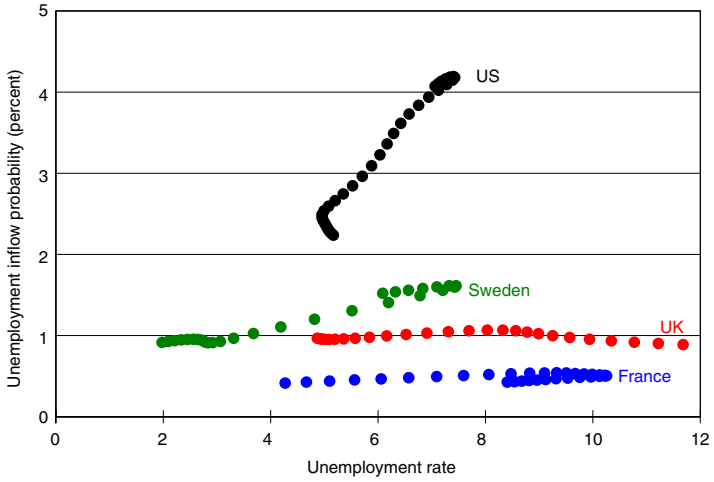


Figure 30 ***Unemployment rate and unemployment inflow probability in France, Sweden, the United Kingdom, and the United States.***

large differences in unemployment inflow probabilities. At an unemployment rate of six percent, the unemployment inflow probability varied by more than a factor of four.

We next consider the issue of how changes in trend unemployment can be decomposed into changes in inflow and outflow probabilities. Elsby et al. (2008) argue that for this purpose one should focus on the log of the inflow and outflow probabilities and we follow this practice. Figures 31 and 32 provide scatter plots for unemployment rates

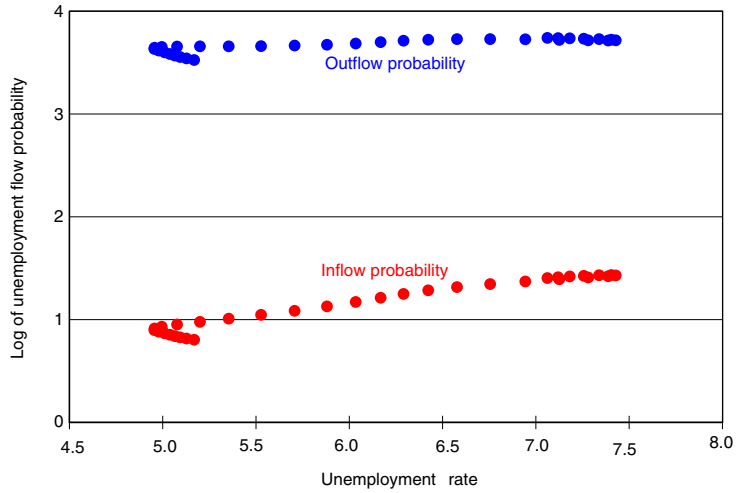


Figure 31 ***Log unemployment inflow and outflow probabilities for the United States.***

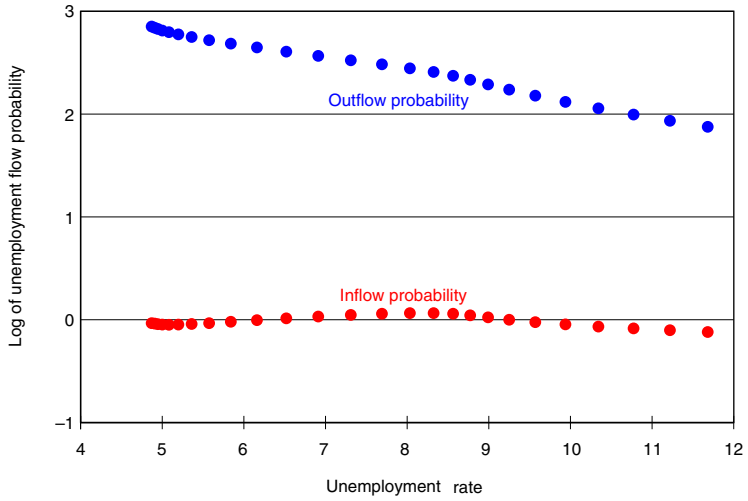


Figure 32 ***Log unemployment inflow and outflow probabilities for the United Kingdom.***

and the log of inflow and outflow probabilities for the United States and the United Kingdom. These two figures are of particular interest because they follow very different patterns. In the United States, increases in trend unemployment were associated with effectively no change in the unemployment outflow probability and an increase in the unemployment inflow probability. That is, changes in trend unemployment were entirely accounted for by changes in unemployment inflows. In contrast, the exact

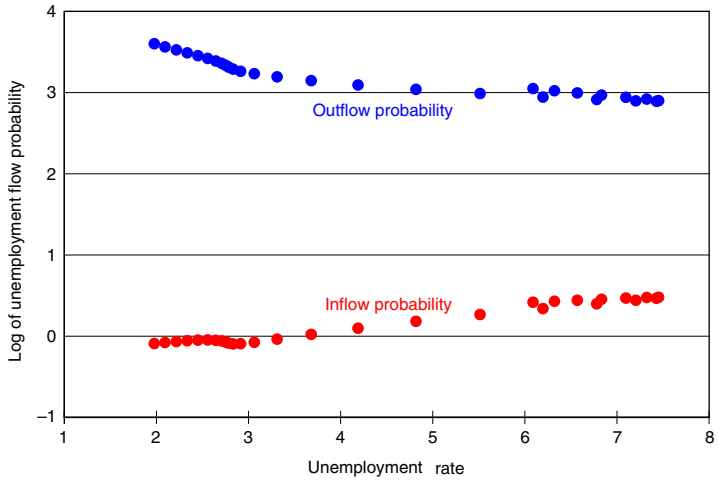


Figure 33 ***Log unemployment inflow and outflow probabilities for Sweden.***

opposite is found in the United Kingdom. Here there was effectively no change in the unemployment inflow probability, implying that changes in trend unemployment were accounted for by changes in unemployment outflow probabilities.

Next we consider the dynamics for Sweden, presented in Fig. 33. In this case, increases in trend unemployment were accounted for by an increase in the unemployment inflow probability and decrease in the outflow probability that were almost equal in magnitude. Recall that the trend unemployment rate followed a single peaked shape, so that there is an interval of unemployment rates that occurred both during the upswing and during the downswing. It is interesting to note that the relationship between the flows and unemployment appears to have been fairly stable over time within Sweden, in the sense that the flow probabilities associated with a given unemployment rate were independent of whether the unemployment rate was increasing or decreasing.

Finally, we consider France in Fig. 34. Similar to the case of Sweden, we see that increases in trend unemployment were accounted for both by a decrease in the outflow probability and an increase in the inflow probability. However, here there was a marked asymmetry between the pattern during the upswing and downswing in (trend) unemployment. Although the points are not labeled by year, the lower portion of the scatter plots for unemployment rates higher than eight percent correspond to the post-1995 period, when trend unemployment was falling. During the pre-1995 period, both factors contributed to the increase in the unemployment rate, although the contribution of the change in the outflow probability was somewhat greater. Indeed, looking more closely, we see that the unemployment inflow probability stopped increasing when unemployment hit approximately eight percent. The subsequent increase in the unemployment rate was accounted for entirely by a decrease in the outflow probability. After

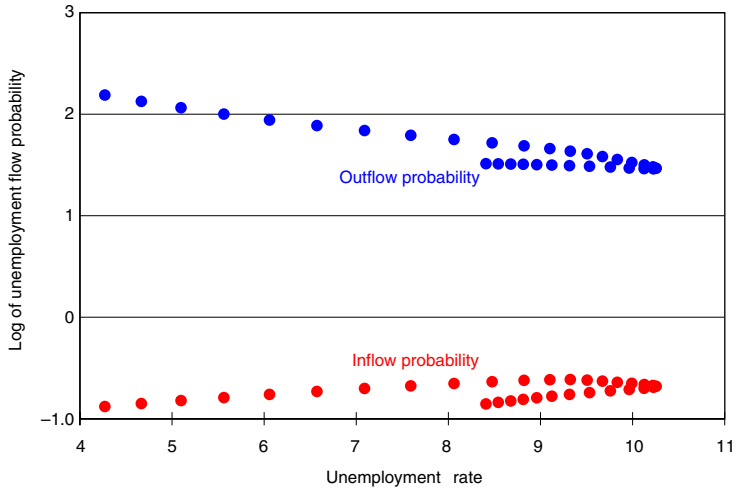


Figure 34 ***Log unemployment inflow and outflow probabilities for France.***

1995, trend unemployment started to fall. This was entirely accounted for by a decrease in the unemployment inflow probability, with no change in the outflow probability.

In summary, and in contrast to business cycle frequencies, there are no strong empirical regularities on worker flows that hold across all OECD countries. There were large differences in worker flows across countries even when unemployment rates were the same. And in some countries inflow probabilities explained most of the trend movements in unemployment, in other countries outflow probabilities were more important, and in still other countries the importance of the two factors varied over time.

2.1.4. Labor wedge

In Section 1.1.6, we analyzed the cyclical properties of the labor wedge and argued that, from the perspective of a standard representative household model, recessions are times when it looks like the tax on labor is high. Prescott (2004) uses the same approach to analyze the trend change in hours worked in the G-7 countries between the early 1970s and early 1990s. He verified that variation in hours worked over time and across countries is associated with variation in the labor wedge, but also argued that the wedge was largely accounted for by measured labor and consumption taxes.

Ohanian et al. (2008) extends this analysis to more countries and a longer time period. While they found that changes in labor and consumption taxes account for a large share of the change in the labor wedge for many countries, there are movements in the labor wedge beyond those that can be explained by measured taxes. In some cases the wedge that remained after accounting for labor taxes was positive, suggesting that individuals are not able to work as much as they would like, while in others it was negative, suggesting that individuals were working more than they would like. Subsequent work

has tried to account for these cases as well. Rogerson (2008) and McDaniel (2009) argue that incorporating trend movements from home production to market work (Aguiar and Hurst, 2007; Ramey and Francis, 2009) helps to decrease the absolute size of the labor wedge in countries such as the United States. And Ragan (2004) and Rogerson (2007) argue that modeling detailed features of government spending in Scandinavia, such as child care subsidies, helps to explain the relatively small labor wedge in those countries.

More generally, the fact that changes in labor and consumption taxes account for a large share of the changes in labor wedge is consistent with a model where individuals are on their labor supply curves, after taking taxes into account. This supports our argument that frictions are not a key part of the story behind the large trend changes in hours worked.[47]

2.2. Theory

This section reviews two very distinct ways that search theory has been used to understand the hump-shaped pattern in trend unemployment and the large differences in worker flows. First, some papers have argued that workers' search decisions play a quantitatively important role in understanding both cross-country differences in worker flows and the different evolutions of aggregate unemployment. For example, if workers choose to look less intensively for jobs and become more selective about which jobs to accept, then holding all else constant, unemployment durations will be longer. Second, other papers have taken advantage of the fact that search creates a tractable framework with match specific rents in order to explore whether wage determination is important in accounting for cross-country unemployment patterns. For example, different labor market policies can influence the manner in which match specific rents are shared and so change the profitability of job creation. This affects unemployment flows and potentially also changes how trend unemployment responds to shocks.

We begin our discussion with a bit of background and context. The initial large and persistent increase in unemployment in some major European countries generated a nascent literature seeking to explain the phenomenon; see Blanchard (2006) for a recent review. Here we sketch a few details. From the outset, theories included two key components: shocks (or driving forces) and propagation mechanisms. Simply put, something must have changed to increase the unemployment rate, and the increase in unemployment reflected the process through which these changes were propagated through the economy. An important early contribution by Bruno and Sachs, summarized in Bruno and Sachs (1985), argued that shocks to oil prices and the slowdown in productivity growth were the key driving forces, and that the failure of real wages to adjust was the key propagation channel. Search theory played very little role in this analysis.

[47] Krusell et al. (2009) show that adding empirically plausible search frictions has virtually no impact on the relationship between taxes and total hours.

A related research effort, which culminated in the first edition of Layard et al. (2005), further developed a framework for analyzing how the interaction of shocks and institutions influenced unemployment. Although this analysis did attribute some role to the search behavior of workers, it did not involve any formal modeling of the search and matching process.[48]

As high unemployment persisted long after the oil shocks had dissipated and beyond the initial decline in productivity growth, researchers looked for other driving forces. Krugman (1994) emphasized broad-based technological change as the shock. The starting point for Krugman's theory was the observed increase in wage dispersion in the United States, particularly between low and high skill workers.[49] Following many others, Krugman attributed this increased dispersion to skill-biased technological change. He posited that this shock should be present in all advanced economies. In the face of this common shock, labor market responses differed across countries because of differences in labor market institutions. Consistent with the earlier literature, Krugman also emphasized wage setting institutions. But whereas the earlier literature had focused on how wage setting institutions affected the change in the overall wage level, Krugman focused on how institutions affected the change in wage dispersion. In the United States, he argued that wages were largely set in competitive markets, so that skill-biased technical change increased wage dispersion but left unemployment relatively unchanged. In contrast, he argued that in many European economies, wage setting institutions did not allow wages to become more spread out. Instead, unemployment increased for less-skilled workers. Again, this explanation did not attribute any role to search.

Although Krugman's theory was intuitively appealing, subsequent work by Card et al. (1999) found little support for the key mechanism in their study of Canada, France and the United States. Still, the dominant approach remains the "shocks-and-institutions hypothesis": different unemployment evolutions are accounted for by a common shock that is propagated differently across countries because of institutional differences (Blanchard and Wolfers, 2000).[50] Numerous studies document differences in a variety of labor market institutions, including labor taxes, employment protection, minimum wages, unemployment benefits, and the nature of wage setting. The dramatic differences in unemployment flows across countries even at the same unemployment rate suggests that these institutions affect labor market outcomes; see Nickell and Layard (1999). A different possibility is that unemployment is changing over time across countries because institutions are changing across countries. But Blanchard and Wolfers (2000) found that this was not supported by panel measures of several labor market institutions. In

[48] Some of this work is summarized in the earlier contributions to this handbook by Johnson and Layard (1986) and Nickell and Layard (1999).

[49] See the previous handbook chapter by Katz and Autor (1999) for an excellent survey of this issue.

[50] A notable exception is Daveri and Tabellini (2000), who emphasize cross-country differences both in driving forces and institutions. More recently Nickell et al. (2005) have also argued that differences in driving forces are important.

particular, while they argued that there has been some change in institutions over time, these changes account for very little of the observed changes in unemployment. On the other hand, they argued that the data does support the view that unemployment changes could be accounted for by common shocks which are propagated differently because of differences in institutions.

While Blanchard and Wolfers (2000) pointed researchers in a particular direction, the paper identified neither the common shock nor the key economic propagation mechanisms. Recent work has sought to isolate the quantitatively important shocks and propagation mechanisms, and in particular the institutional features that impact the propagation of these key shocks. Not surprisingly, since search models have become the dominant models for analyses of unemployment, most of the recent work in this literature has taken place in theoretical models that feature search. Prominent examples include Bertola and Ichino (1995), Ljungqvist and Sargent (1998, 2004, 2007), Marimon and Zilibotti (1999), Mortensen and Pissarides (1999c); Pissarides (2007), and Hornstein et al. (2007). Each of the above papers is also implicitly a theory of cross-country differences in unemployment flows. Several other papers have used search theory to explore unemployment flows without necessarily addressing the issue of how unemployment has evolved over time; see Bertola and Rogerson (1997), Garibaldi (1998), Blanchard and Portugal (2001), Kugler and Saint-Paul (2004), and Pries and Rogerson (2005).

It is beyond the scope of this chapter to review this large literature. Instead, we focus on two examples which illustrate the two roles that search theory has played. Ljungqvist and Sargent (1998 hereafter referred to as LS) focus on the role of worker choices regarding search. Hornstein et al. (2007 hereafter HKV) focus on how wage determination affects the profitability of job creation. A closely related point, also of interest in a broader sense, is that the two papers propose fundamentally different views of the underlying economic forces that have led to higher unemployment. While LS focus on the role of worker choices, with firms playing a passive role, HKV focus on the choices of firms, with workers playing a passive role.

2.2.1. Ljungqvist and Sargent (1998)

LS quantify one intuitively plausible version of the shocks-and-institutions hypothesis. Their common shock is an increase in "turbulence." In their language, when a worker is laid off from her job in a turbulent economy, her human capital often becomes less valuable or depreciates. A simple example serves to illustrate what they have in mind. In one scenario, a worker in an auto assembly plant loses her job due to lower demand for the particular model produced by that plant, but then finds employment at a different auto assembly plant that is hiring workers because of increased demand for its particular model. Although the worker may suffer a spell of unemployment, she suffers no long term wage loss. In a second scenario, there is a permanent decrease in employment in all domestic auto assembly plants, perhaps because of labor-saving technological change

or competition from lower-cost assemblers in other countries. In this case, the laid-off worker not only loses her job and experiences a spell of unemployment, but also suffers a drop in expected future wages since the demand for her skills has fallen. LS argue that this second scenario became more prevalent in all OECD countries starting some time in the late 1970s and early 1980s.

LS interact this common shock with differences across countries in social insurance schemes for dealing with workers who suffer job losses. Although they label the policies "unemployment insurance," they interpret them broadly to include other programs for displaced workers, such as disability and early retirement.[51] In their analysis, LS focus on two different institutional regimes. In one, which they label laissez-faire, there is no social insurance at all. In the other, which they label the welfare state, a worker who loses her job can collect a transfer payment equal to 70 percent of her pre-layoff wage as long as she remains jobless, subject only to the proviso that she cannot reject a job that offers at least as high of a wage as she earned before the layoff.

Frictionless model Our goal is to assess whether search frictions per se improve our understanding of low frequency changes in labor market outcomes, and so we compare frictional and frictionless version of the LS model. We start with the simpler frictionless model. There is a unit mass of workers, each of whom seeks to maximize a utility function of the form:

$$\sum_{t=0}^{\infty}(\beta(1-\alpha))^t y_t,$$

where β is the discount factor, α is the probability of death (assumed constant), and y_t is after-tax income in period t. When an individual dies, she is replaced by a newborn individual. Note that individuals do not value leisure. This means that, even though the analysis emphasizes the importance of labor supply responses, these responses are not based on substitution between leisure and consumption.

The key features of the LS model are learning-by-doing when working and skill depreciation when not working. Let h denote the human capital or skill level of a worker. A newborn worker starts with some minimal level of human capital, which we normalize to 1. Subsequently, there are three transition functions that describe how human capital evolves. Let $\mu_e(h, h')$ be the probability that a worker who remains employed goes from human capital h to human capital h'. Similarly, let $\mu_n(h, h')$ be the probability that a worker who is currently not employed goes from human capital h to human capital h'. Finally, to capture turbulence as defined above, let $\mu_l(h, h')$ be the probability that an employed worker who gets laid off goes from human capital h to human capital h'. In the economy without turbulence, there is no loss in human capital upon layoff.

[51] One issue in terms of connecting this analysis with the data is that individuals in these programs will typically not be classified as unemployed.

LS assume a finite set of levels of human capital and parameterize the stochastic human capital accumulation process as follows: when employed, human capital either stays the same or improves by one level. When not employed, human capital either stays the same or deteriorates by one level. It follows that human capital for a continuously employed (non-employed) individual is weakly increasing (decreasing). Finally, when a worker with human capital h is laid off in the turbulent economy, there is some probability that her human capital remains unchanged, but with the remaining probability the worker receives a draw from a distribution with support on $[1, h]$.

There is an unlimited supply of identical jobs, each of which offers a wage w per unit of human capital. That is, a worker with human capital h earns wh if she works. We introduce the base wage w to facilitate the comparison with the search theoretic version of LS, but it plays no role in the frictionless model. When a worker is laid off, she can immediately move to a new job if she wants. This means that, while layoffs can have a negative consequence for human capital, they do not affect the opportunity to work.

The final feature of the economic environment is a balanced-budget unemployment insurance system, financed by a proportional tax on income. The laissez-faire economy has no unemployment insurance, while the welfare state economy pays a laid-off worker 70 percent of her pre-layoff earnings until the worker opts to return to work.

LS study four steady state outcomes: the laissez-faire and welfare state, each with and without turbulence. First consider the laissez-faire economy without turbulence, so human capital is unchanged following a layoff. In equilibrium, all individuals work until they die. This follows immediately from the fact that individuals do not value leisure and always have access to a job that offers positive earnings. The equilibrium of the welfare state without turbulence is the same. Although in this economy a laid-off individual has access to life-long unemployment benefits equal to 70 percent of her pre-layoff earnings, she can always find a job that offers 100 percent. Since she does not value leisure, it is optimal to work. The basic message is that, in some environments, large differences in unemployment insurance do not affect equilibrium employment.

Next we repeat this analysis when there is turbulence in the economy. The equilibrium employment decisions in the laissez-faire economy are unchanged. Even though a laid off worker may experience a large negative shock to her human capital, and thereby face lower earnings prospects, it is still optimal for her to work, since leisure has no value. But the equilibrium outcome in the welfare state economy is potentially affected. If a worker experiences more than a 30 percent reduction in her human capital, then the one period return to working is now lower than the one period return to collecting unemployment insurance. Of course, working allows the individual to accumulate human capital, and so it may still be optimal for her to work. But in general, a laid-off worker will stop working if the shock is sufficiently bad. If she does not work in the period following the layoff, then her human capital starts to depreciate. Since unemployment benefits last forever, she will never find it optimal to return to work.

This discussion highlights two key points. First, the impact of unemployment insurance, and other social insurance programs more generally, on aggregate employment depends on the underlying economic environment. In the absence of turbulence, even very generous programs need not affect aggregate employment. Second, the LS model offers an example in which a common shock to economies that have different labor market institutions can be propagated very differently, with very different effects on aggregate employment. This holds even in the absence of frictions.

Model with search frictions We now extend the model to allow for search frictions in the spirit of Lippman and McCall (1976), as in the original LS article. A worker without a job makes a decision regarding search intensity (s) that influences the probability of receiving a wage offer w, drawn from the cumulative distribution function F with bounded support. The probability that a worker receives an offer is an increasing function $\pi(s)$, while the utility cost of searching while unemployed is an increasing function $c(s)$. There is no on-the-job search. Note that a worker's income is the product of her wage and human capital, wh. The wage stays constant on the job, but her human capital may increase. In the welfare state, unemployment benefits are tied to past income.

Search frictions affect model outcomes both through search intensity and through the reservation wage. To understand how, note that in a standard worker search problem without human capital, unemployment insurance reduces search intensity and raises the reservation wage, both of which lead to longer unemployment durations and hence higher unemployment. These forces are present even in this more complicated environment.

This observation has two implications. First, it implies that even in the absence of turbulence, the welfare state economy has higher unemployment than the laissez-faire economy and that this higher unemployment is driven by differences in duration, although the effects are quantitatively modest.

Second, the behavioral response through search intensity and reservation wages amplifies the effect of turbulence. Recall that in the frictionless model, a worker who experiences a sufficiently large shock when laid off becomes non-employed forever. With search frictions and skill depreciation while nonemployed, human capital continues to fall after the initial layoff. This means that, even if the initial shock did not leave the worker preferring nonemployment to working, she may slip into this absorbing state during a long unemployment spell. Indeed, if she approaches a point at which she prefers unemployment to work at any wage, her search intensity falls and reservation wage rises, so the job finding probability falls smoothly to zero.

The discussion so far has been entirely qualitative. A key contribution of LS is to assess the quantitative importance of these factors. To do this, LS compute the steady state equilibria for each of the two UI regimes both with and without turbulence. The no-turbulence laissez-faire economy is parameterized so as to match several features of the United States economy in the 1970s. LS assume that all economies share the

same preferences and technologies, and that the only difference is the unemployment insurance system.

In the absence of turbulence LS find that both economies have similar steady state unemployment rates, although long duration unemployment is more prevalent in the welfare state economy with the UI system. This comparison is important, since as we noted previously, unemployment duration was higher in France and Sweden than in the United States, even when unemployment rates were similar.[52] But when they introduce turbulence, they find that the welfare state moves to a steady state with much higher unemployment, whereas unemployment in the laissez-faire economy changes relatively little. In addition, all of the increase in unemployment is accounted for by increased duration of unemployment. Based on these findings, LS conclude that an increase in turbulence combined with different unemployment insurance systems is quantitatively important in accounting for the variation in unemployment rate evolutions across countries.

As we noted before, the higher unemployment that appears in the welfare state in turbulent times is purely the result of worker choices. Indeed, there are no firms in the model. In later work, Ljungqvist and Sargent (2004) extend the framework to an environment with endogenous job creation and show that their results are basically unchanged. For our purposes, the most important observation is that one can in principal obtain similar results even without search frictions.

2.2.2. Hornstein et al. (2007)

HKV propose an alternative version of the shocks-and-institutions hypothesis. In this case, the shock is an acceleration in capital embodied technological change, consistent with the acceleration of the decline in equipment prices from the 1960s to the 1990s documented by Gordon (1990) and Cummins and Violante (2002).[53] As in LS, HKV examine the effect of different unemployment benefit systems, but also broaden the set of relevant institutions to include labor income taxes and firing taxes. They find that the effect of an increase in the pace of technological change depends on how wages behave, which in turns depends on the amount of rents in the initial equilibrium, itself a function of these institutions.

Frictionless model We again start with a version of the model that does not include search so as to better understand what role search plays in HKV's analysis. There is a unit mass of identical workers, each with preferences

$$\int_0^\infty \mathrm{e}^{-\rho t}(c(t) - \gamma n(t))\mathrm{d}t,$$

[52] In fact, unemployment was actually lower in Europe than in the United States in 1970. Ljungqvist and Sargent (2008) show how to modify their analysis to capture this feature.

[53] Earlier work on technological change in the context of search models includes Aghion and Howitt (1994) and Mortensen and Pissarides (1999c).

where $c(t)$ is consumption at time t, $n(t) \in \{0, 1\}$ is labor supply at time t, and ρ is the discount rate. The unit of production is a matched worker-machine pair, so the microeconomic production technology is Leontief in machines and labor. Machines are indexed by age and a matched worker-machine pair with age of machine a produces output e^{-ga}. Output of a new machine is normalized to one. The parameter g embodies two forces: depreciation and capital embodied technological change. Relative to the newest vintage, both of these forces imply a negative effect of age on productivity. A new machine costs I units of output. There is an unbounded set of potential entrants. A key assumption is that machines cannot be upgraded. The scrap value of a machine is normalized to zero.

Consider the steady state competitive equilibrium in this economy. There are two markets, one for output and one for labor. Let output be the numeraire and let w denote the wage.[54] Because this model features linear technology and preferences, equilibrium will generically entail either everyone working or no one working. The only interesting case is the one in which everyone works, so in what follows we will assume that parameters are such that this is true in equilibrium.

The only real decision in this economy is the age at which a machine gets scrapped. Given a wage of w, a machine is scrapped at the age a solving $w = e^{-ga}$, since this is the point at which it is no longer profitable to operate the machine. Given that there is an unbounded set of potential entrants, the equilibrium w is the one at which the present discounted value of profit from entering is zero. This implies

$$\int_0^{-\frac{\log w}{g}} e^{-\rho a}(e^{-ga} - w)\mathrm{d}a = I,$$

so the present value of profits during the productive life of the machine just covers the initial cost of the machine. This equation uniquely determines w.

Now consider what happens to the steady state equilibrium of this economy if there is an increase in the rate of obsolescence, i.e., an increase in g. Implicitly differentiating the above expression shows that w must fall. The basic intuition is that if a machine becomes obsolete more rapidly then it will be used for a shorter period of time, implying that wages must decrease in order to maintain the profitability of investment. It also follows that the scrapping age will decrease. The same must be true if there is an increase in the rate of technological change: the wage relative to the productivity of the frontier technology must fall, and the scrapping age decreases as well. In the economy without search frictions, this has no consequences for unemployment or employment. Workers move directly from a machine that is being scrapped to a new machine. There will, however, be more reallocation of workers when g is larger. But the key point is that in order to

[54] We follow HKV in measuring everything relative to the frontier technology so as to make the economy stationary. This means that in our exposition, the wage is constant relative to the frontier, and so is actually growing over time if there is capital embodied technological change.

maintain equilibrium in the labor market, increases in the pace of technological change necessitate a decrease in wages relative to the productivity of the frontier technology.

Model with search frictions Now we return to the full HKV model with search frictions. Similar to the model in Section 1.2, frictions are modeled through the use of a constant returns to scale matching function, but here the inputs are unmatched workers and machines. There are no search costs, only the upfront cost of a new machine. Workers separate from matches for exogenous reasons at rate x. The machine still exists after a separation of this sort and will look for a new worker.

Wage determination is the key to HKV's analysis. The presence of search frictions gives rise to match specific rents and there are many ways that these can be divided. Following the approach pioneered by Pissarides (1985) and Mortensen and Pissarides (1994), HKV assume that wages satisfy the Nash bargaining solution. This gives rise to a wage function $w(a)$. In particular, and in contrast to the frictionless model, workers who work with machines of different vintages are paid different amounts, since the rents associated with a given match depend on the age of the machine.

As before, free entry ensures that in equilibrium firms earn zero expected profits from the purchase of a new machine. But two endogenous factors now influence profits: the wage function and the fraction of time that the machine will be idle. Increased idleness and higher wages both reduce profits. Zero profit can be achieved either with low idleness and high wages or high idleness and low wages, or combinations in between. Note that, with a constant returns to scale matching function $m(u, v)$, a decrease in the idleness of machines can occur only if unemployed workers find vacant machines at a lower rate, thereby increasing unemployment duration. Thus equilibrium requires either higher unemployment and higher wages or lower unemployment and lower wages. What happens depends on the wage function $w(a)$.

With Nash bargaining, the response of wages to an increase in the rate of skill-biased technical change g is influenced by various labor market policies. Intuitively, wages can be thought of as having two components, one associated with the worker's outside option and the other with the worker's share of the match specific rents. As the match specific rents become smaller, rent-sharing represents a smaller fraction of the total wage. As a result, the total wage responds less to g and frictions respond more. Labor income taxes and unemployment benefits tend to reduce the surplus. To the extent that these policies are more prevalent in Europe than the United States, there is less wage adjustment in Europe and so more adjustment in the amount of time that workers spend idle.

Once again, a key issue is the quantitative importance of this mechanism. HKV calibrate the model to match features of the United States economy in the early 1970s. They then solve for the steady state that corresponds to policy settings that represent a "typical" European economy, assuming that Europe is identical to the United States in all non-policy factors except x, the rate at which workers separate from jobs for exogenous

reasons.[55] They then increase the pace of technological change to the rate observed in the 1990s and solve for the steady state equilibrium for both the United States and European economies. Their key finding is that unemployment is much more responsive to the pace of technological change in Europe than in the United States. The outflow from unemployment falls, due to the decrease in investment in new machines, and the inflow rate increases, due to the reduction in the scrapping age. Quantitatively, HKV find that the reduction in outflows accounts for most of the increase in unemployment.

As noted earlier, and in contrast to the analysis in LS, increased unemployment in this model is attributed entirely to firms' actions. Workers are completely passive, accepting any job that they are offered. Unemployment increases primarily because firms create fewer new machines and so post fewer vacancies. But the main role played by search is that it gives rise to match-specific rents. Differences in how these rents are divided are important for the behavior of aggregate unemployment.

2.2.3. Discussion

The main goal of this section has been to highlight how search theory has been used to understand cross-country differences in labor market outcomes. Although we discussed two papers in more detail, we view these as representative of a larger literature that has sought to flesh out the shocks-and-institutions hypothesis. In closing, we discuss how well these papers account for the key features of the data that we have previously described. We focus on four dimensions.

First, both models look only at the movement of workers between employment and unemployment; neither has a participation margin or an hours margin. We previously documented that the unemployment margin accounts for a relatively small fraction of the overall changes in total hours, which raises the issue of whether it is appropriate to start with a framework that abstracts from the quantitatively more important margins. More generally, given that something must be generating these large changes in hours and participation, it is unclear if it is reasonable to ignore these changes when thinking about unemployment.

Second, both models solve for two steady state equilibria, one that corresponds to the "earlier" period and the other that corresponds to the "later" period. But as we documented previously, it does not look like Europe moved from a low unemployment steady state to a high unemployment steady state. Instead, it seems more promising to think of a shock that led to a long-lived but at least partially temporary increase in unemployment. One possibility is that the shocks themselves are long-lived but temporary, but at a minimum this points to the need for better measurement of the key forcing processes. For example, LS provide indirect evidence on the increase in turbulence, based

[55] As noted above in the discussion of the LS model, a key fact is that unemployment in Europe was not always higher than in the United States, despite the fact that European institutions have remained relatively constant. The lower value of x in Europe is critical in allowing for the HKV model to be consistent with this fact.

on work by Gottschalk and Moffitt (1994) that decomposes earnings in the 1970s and 1980s into transitory and permanent components. LS show that their calibrated model can replicate the key features of the change in the decomposition in the United States. While this decomposition provides some useful information, Gottschalk and Moffitt (1994) do not directly measure the wage losses associated with layoffs. Nor has anyone checked whether the changes in the decomposition were reversed after 1995, when unemployment started to fall throughout the OECD. Finally, both LS and HKV impose that the shocks are the same across countries, without offering direct evidence on this point. That is, they do not directly test the shocks-and-institutions hypothesis.

Third, both models contrast outcomes in two economies, the United States and a "typical" European economy. But a key feature of our earlier data analysis is that there is a large amount of heterogeneity in unemployment dynamics even across European economies, both in terms of the magnitude of the increase and in terms of the timing of the peak. It is also well known that there is substantial heterogeneity in institutions and policies across economies. Accounting for the heterogeneity in outcomes across countries is an important test for any theory of unemployment dynamics.

Fourth, both papers focus entirely on aggregate outcomes. There is potentially a lot of useful information in the disaggregated data that may help to distinguish competing theories. Although we have not presented any detailed information about the patterns of labor market outcomes in disaggregated data, we note two findings here: unemployment differences are particularly pronounced among younger workers, while employment differences are particularly pronounced both for younger and older (but not prime-aged) workers. More generally, incorporating differences in age, gender, and skill level may be useful for distinguishing theories of labor market outcomes.[56]

Finally, we want to address the role of search in these two papers. While LS emphasize workers' choices about search intensity and reservation wages, they do not show that search frictions are important for propagating turbulence. To our knowledge, this question remains unanswered. More generally, whether search intensity amplifies or mutes the response of the economy to shocks may depend on the nature of the shock. For example, the unemployment consequences of shocks that require reallocating workers across sectors may be muted by the possibility of searching more intensively.

The analysis of HKV raises several interesting questions. First, we have emphasized that differences in rent-sharing is a key mechanism in their model. While search models represent a tractable framework that gives rise to match specific rents, we previously noted that other mechanisms may also give rise to match-specific rents. We know of no attempt to measure whether search frictions are an important source of match-specific rents, nor of any work that has examined whether the source of rents is important for labor market outcomes.

[56] Kitao et al. (2009) is a recent example that moves in this direction.

Second, note that HKV shares a common feature with Bruno and Sachs (1985) and Krugman (1994), in that all three papers emphasize that unemployment increases when wage setting institutions prevent wages from responding sufficiently following a particular shock. A key difference is that HKV assume that wage setting institutions are fundamentally the same across economies and over time, represented by a Nash bargaining solution with the same bargaining weights in each country and at each point in time. Wage outcomes differ because of the interaction between the wage setting institution and other labor market policies. Two interesting and open questions are the extent to which different economies in fact have different wage setting processes and, if they do, how to model those differences. This is of course not inconsistent with search theory. Indeed, one nice feature of the Mortensen and Pissarides (1994) framework is that it can easily accommodate a wide variety of specifications for wage setting.[57] Nonetheless, little work has sought to assess the extent to which different countries should be modeled as having different wage setting processes. Recent work that leads in this direction includes Mortensen (2003), who asks which of two wage setting specifications better fits Danish microeconomic data, and Hall and Krueger (2008), who use survey evidence to assess the relevance of different wage setting procedures in the United States economy.

3. CONCLUSION

Our objective in this chapter has been to explore how the explicit introduction of search frictions into otherwise standard macroeconomic models affects our understanding of aggregate labor market outcomes in two different contexts. In our analysis of business cycles, we found that the search framework is useful for interpreting facts about unemployment and labor market flows. But we also found that search frictions tend to dampen fluctuations in output and employment without significantly increasing their persistence. Moreover, by dampening employment fluctuations, search frictions cause a counterfactual procyclical labor wedge. In addition, using search theory did not lead us to introduce any important new shocks into the neoclassical growth model. On all of these counts, search behavior itself does not seem intrinsically important for business cycle analysis. On the other hand, search provides a promising environment for studying the implications of alternative wage setting mechanisms, which many authors have argued are important at business cycle frequencies. We remain hopeful that search will be an important component in understanding business cycle fluctuations. In particular, we think that the ability of search models to connect with data on job and worker flows will help to discriminate between alternative theories of wage setting behavior.

[57] See the handbook chapter by Mortensen and Pissarides (1999c) for details on how different models of wage determination can be embedded in search-theoretic models of the labor market.

We also found some striking low-frequency patterns in unemployment and labor market flows. For example, trend unemployment initially rose across the OECD but has since fallen in almost every country. And there is substantial heterogeneity in the relationship between unemployment and worker flows across countries and over time. Search is useful for interpreting these facts. But unlike in the business cycle analysis, changes in trend unemployment are typically much smaller than changes in hours per worker and labor force participation. To the extent that search models lead one to focus only on unemployment, we feel that the emphasis is misplaced. Still, many recent explorations of the shocks-and-institutions hypothesis have taken place in models that feature search. While the role of search frictions per se in these models is unclear, we are also hopeful that these models will prove useful for understanding the extent to which differences in wage setting institutions as well as other labor market institutions and policies are an important cause of differences in labor market outcomes across countries.

In interpreting our conclusions it is important to emphasize that we have focused throughout on the role of search in *macroeconomic* models of the labor market. But we would be remiss not to mention at least briefly the important role that search frictions play in four distinct branches of the microeconomic literature on labor market outcomes.[58]

First, search theory has served as the foundation for the analysis of optimal unemployment insurance. Shavell and Weiss (1979) explored the optimal provision of unemployment insurance to a worker who must both choose how intensively to search and what wages to accept. Most of the subsequent literature has focused on the search intensity margin (e.g., Hopenhayn and Nicolini, 1997; Chetty, 2008), although a few papers have looked at how unemployment insurance affects reservation wages (Hansen and Imrohoroğlu, 1992; Shimer and Werning, 2008). Search is fundamental to these papers, since each needs a model with unemployment and a moral hazard problem in an environment with idiosyncratic risk. More generally, models that feature search have been used to analyze many labor market policies. Ljungqvist (2002), for example, shows that a model with search frictions has distinctive implications for the effects of employment protection on aggregate employment.

Second, search models have been useful in accounting for worker flows. There is a long tradition of using single agent search theory to account for individual level data on unemployment spells and wages.[59] Search behavior plays an essential role in this literature. More recently, equilibrium search models have been used to study data on turnover and wage dynamics. Many are based on the Burdett and Mortensen (1998) model of on-the-job search; prominent examples include Postel-Vinay and Robin (2002),

[58] Search frictions have also been used to model monetary exchange, housing, marriage, and over-the-counter asset markets, among other topics.

[59] Early examples include Flinn and Heckman (1982) and Wolpin (1987). See also the discussions in Mortensen (1987) and Eckstein and Van den Berg (2007) and the references contained therein.

Mortensen (2005), and Cahuc et al. (2006). These papers take advantage of the availability of large administrative matched worker-firm panel data sets to estimate the structural parameters of the model and use the data to test alternative theories of wage determination. See Lentz and Mortensen (2010) for a recent survey of this literature.

Third, a number of authors have argued that search frictions may play an important role in understanding the evolution of wage inequality. For example, Acemoglu (1999) presents a model where an increase in the supply of skilled workers may change the composition of jobs from a "pooling" equilibrium, in which firms facing search frictions create jobs suitable for all workers, to a "separating" equilibrium, in which firms create different types of jobs for different workers, leading to an increase in between-group wage inequality. Other authors have explored the extent to which search models, by permitting violations of the law-of-one-price, can generate within-group wage inequality. Hornstein et al. (2009) argue that within a broad class of search models, the possibility of an economically significant amount of within-group inequality is limited by the possibility of waiting for better job opportunities. On the other hand, Marimon and Zilibotti (1999) show how the interaction between search frictions and unemployment benefits can generate substantial inequality across ex ante identical workers. Shimer and Smith (2000), Shi (2002), and Shimer (2005a) develop theoretical search models with heterogeneous workers and firms that make predictions for both within- and between-group inequality. These models have been explored empirically, for example by Abowd et al. (2004) and Lopes de Melo (2009).

Finally, a few papers have used search frictions to address existing issues in the labor contracting literature. In particular, building on Gale (1996) and Inderst and Wambach (2001, 2002), Guerrieri et al. (2009) show that search can naturally resolve problems related to the existence and uniqueness of equilibrium in adverse selection models. Their applications include a labor market rat race, where high productivity workers agree to an inefficiently high level of hours and employment in order to separate themselves from less productive workers, and the absence of private insurance against the risk of a layoff, where high productivity workers are willing to work without insurance in order to indicate that they are not concerned with this eventuality. More generally, one might expect search frictions to interact with standard issues that arise in labor contracting.

In concluding, it is useful to return to our discussion in the introduction regarding the three different ways that search might matter in macroeconomic models. At this point, we do not see much evidence that search behavior per se is of first order importance in understanding aggregate outcomes in either of the contexts that we considered. However, adding search to otherwise standard macroeconomic models definitely expands the ability of these models to connect with various pieces of empirical evidence. And models that feature search do create a useful framework in which to consider various wage setting mechanisms. Further clarifying the role of search in assessing substantive issues involving the aggregate labor market is an important task for future research.

REFERENCES

Abowd, John M., Kramarz, Francis, 1999. The analysis of labor markets using matched employer-employee data. In: Ashenfelter, Orley, Card, David E. (Eds.), Handbook of Labor Economics, vol. 3. Elsevier, pp. 2629–2710 (Chapter 40).

Abowd, John M., Kramarz, Francis, Lengermann, Paul, Pérez-Duarte, Sébastien, 2004. Are Good Workers Employed by Good Firms? A Test of a Simple Assortative Matching Model for France and the United States. Discussion paper, Cornell University mimeo.

Abowd, John M., Zellner, Arnold, 1985. Estimating gross labor-force flows. Journal of Business & Economic Statistics 3 (3), 254–283.

Abraham, Katharine, 1987. Help-wanted advertising, job vacancies, and unemployment. Brookings Papers on Economic Activity 1, 207–243.

Abraham, Katharine, Shimer, Robert, 2001. Changes in unemployment duration and labor-force attachment. In: Krueger, Alan, Solow, Robert (Eds.), The Roaring Nineties. Russell Sage Foundation and Century Foundation, pp. 367–420.

Acemoglu, Daron, 1999. Changes in unemployment and wage inequality: an alternative theory and some evidence. American Economic Review 89 (5), 1259–1278.

Acemoglu, Daron, Shimer, Robert, 1999. Efficient unemployment insurance. Journal of Political Economy 107 (5), 893–928.

Aghion, Phillipe, Howitt, Peter, 1994. Growth and unemployment. The Review of Economic Studies 61 (3), 477–494.

Aguiar, Mark, Hurst, Erik, 2007. Measuring trends in leisure: the allocation of time over five decades. Quarterly Journal of Economics 122 (3), 969–1006.

Alvarez, Fernando, Veracierto, Marcelo, 1999. Labor-market policies in an equilibrium search model. NBER Macroeconomics Annual 14, 265–304.

Alvarez, Fernando, Veracierto, Marcelo, 2001. Severance payments in an economy with frictions. Journal of Monetary Economics 47 (3), 477–498.

Andolfatto, David, 1996. Business cycles and labor-market search. American Economic Review 86 (1), 112–132.

Attanasio, Orazio, Low, Hamish, Sanchez-Marcos, Virginia, 2008. Explaining changes in female labor supply in a life-cycle model. American Economic Review 98 (4), 1517–1552.

Azariadis, Costas, 1975. Implicit contracts and underemployment equilibria. Journal of Political Economy 83 (6), 1183–1202.

Baily, Martin N., 1974. Wages and employment under uncertain demand. Review of Economic Studies 41 (1), 37–50.

Barro, Robert J., 1977. Long-term contracting, sticky prices, and monetary policy. Journal of Monetary Economics 3 (3), 305–316.

Bertola, Giuseppe, Ichino, Andrea, 1995. Wage inequality and unemployment: United States vs. Europe. NBER Macroeconomics Annual 10, 13–54.

Bertola, Giuseppe, Rogerson, Richard, 1997. Institutions and labor reallocation. European Economic Review 41 (6), 1147–1171.

Bewley, Truman F., 2002. Why wages don't fall during a recession. Harvard University Press.

Blanchard, Olivier, 2006. European unemployment: the evolution of facts and ideas. Economic Policy 21 (45), 5–59.

Blanchard, Olivier, Diamond, Peter, 1989. The Beveridge curve. Brookings Papers on Economic Activity 1, 1–76.

Blanchard, Olivier, Diamond, Peter, 1990. The cyclical behavior of the gross flows of US workers. Brookings Papers on Economic Activity 2, 85–143.

Blanchard, Olivier, Galí, Jordi, 2007. Real wage rigidities and the new Keynesian model. Journal of Money, Credit and Banking 39 (S1), 35–65.

Blanchard, Olivier, Galí, Jordi 2008. Labor markets and monetary policy: a new-Keynesian model with unemployment. NBER Working Paper 13897.

Blanchard, Olivier, Portugal, Pedro, 2001. What hides behind an unemployment rate: comparing portuguese and US labor markets. American Economic Review 91 (1), 187–207.

Blanchard, Olivier, Wolfers, Justin, 2000. The role of shocks and institutions in the rise of European unemployment: the aggregate evidence. Economic Journal 110 (462), 1–33.
Bruno, Michael, Sachs, Jeffrey D., 1985. Economics of worldwide stagflation. Harvard University Press.
Burda, Michael, Wyplosz, Charles, 1994. Gross worker and job flows in Europe. European Economic Review 38 (6), 1287–1315.
Burdett, Kenneth, Mortensen, Dale T., 1998. Wage differentials, employer size, and unemployment. International Economic Review 39 (2), 257–273.
Burdett, Kenneth, Shi, Shouyong, Wright, Randall, 2001. Pricing and matching with frictions. Journal of Political Economy 109 (5), 1060–1085.
Caballero, Ricardo J., Hammour, Mohamad L., 1996. On the timing and efficiency of creative destruction. Quarterly Journal of Economics 111 (3), 805–852.
Cahuc, Pierre, Postel-Vinay, Fabien, Robin, Jean-Marc, 2006. Wage bargaining with on-the-job search: theory and evidence. Econometrica 74 (2), 323–364.
Card, David, Kramarz, Francis, Lemieux, Thomas, 1999. Changes in the relative structure of wages and employment: a comparison of the United States, Canada, and France. Canadian Journal of Economics 32 (4), 843–877.
Chang, Yongsung, Kim, Sun-Bin, 2007. Heterogeneity and aggregation: implications for labor-market fluctuations. American Economic Review 97 (5), 1939–1956.
Chari, V.V., Kehoe, Patrick J., McGrattan, Ellen R., 2007. Business cycle accounting. Econometrica 75 (3), 781–836.
Chetty, Raj, 2008. Moral hazard versus liquidity and optimal unemployment insurance. Journal of Political Economy 116 (2), 173–234.
Cociuba, Simona E., Prescott, Edward C., Ueberfeldt, Alexander, 2009. US hours and productivity behavior using cps hours worked data: 1959-I to 2009-II. Manuscript, Federal Reserve Bank of Dallas.
Cogley, Timothy, Nason, James M., 1995. Output dynamics in real-business-cycle models. American Economic Review 85 (3), 492–511.
Cooley, Thomas F., Prescott, Edward C., 1995. Economic growth and business cycles. In: Cooley, Thomas F. (Ed.), Frontiers of Business Cycle Research. Princeton University Press, pp. 1–38.
Costain, James S., Jansen, Marcel, 2009. Employment fluctuations with downward wage rigidity: the role of moral hazard. IZA Discussion Paper 4344.
Costain, James S., Reiter, Michael, 2008. Business cycles, unemployment insurance, and the calibration of matching models. Journal of Economic Dynamics and Control 32 (4), 1120–1155.
Cummins, Jason G., Violante, Giovanni L., 2002. Investment-specific technical change in the United States (1947–2000): measurement and macroeconomic consequences. Review of Economic Dynamics 5 (2), 243–284.
Daveri, Francesco, Tabellini, Guido, 2000. Unemployment, growth and taxation in industrial countries. Economic Policy 15 (3), 49–104.
Davis, Steven J., Jason Faberman, R., Haltiwanger, John C., Rucker, Ian, 2008. Adjusted estimates of worker flows and job openings in JOLTS. NBER Working Ppaer 14137.
Davis, Steven J., Haltiwanger, John C., 1999. Gross job flows. In: Ashenfelter, Orley, Card, David E. (Eds.), Handbook of Labor Economics, vol. 3. Elsevier, pp. 2711–2805 (Chapter 41).
Diamond, Peter A., 1982a. Aggregate demand management in search equilibrium. Journal of Political Economy 90 (5), 881–894.
Diamond, Peter A., 1982b. Wage determination and efficiency in search equilibrium. The Review of Economic Studies 49 (2), 217–227.
Eckstein, Zvi, Van den Berg, Gerard J., 2007. Empirical labor search: a survey. Journal of Econometrics 136 (2), 531–564.
Elsby, Michael W.L., Hobijn, Bart, Şahin, Ayşegül, 2008, Unemployment dynamics in the OECD. NBER Working Paper 14617.
Elsby, Michael W.L., Michaels, Ryan, Solon, Gary, 2009. The ins and outs of cyclical unemployment. American Economic Journal: Macroeconomics 1 (1), 84–110.
Fallick, Bruce, Fleischman, Charles, 2004. Employer-to-Employer Flows in the US Labor Market: The Complete Picture of Gross Worker Flows. Federal Reserve Board, Finance and Economics Discussion Series Working Paper 2004-34.

Flinn, Christopher, Heckman, James J., 1982. New methods for analyzing structural labor force dynamics. Journal of Econometrics 18 (1), 115–168.
Frazis, Harley J., Robison, Edwin L., Evans, Thomas D., Duff, Martha A., 2005. Estimating gross flows consistent with stocks in the CPS. Monthly Labor Review 128 (9), 3–9.
Fujita, Shigeru, Ramey, Garey, 2009. The cyclicality of separation and job finding rates. International Economic Review 50 (2), 415–430.
Gale, Douglas, 1996. Equilibria and Pareto optima of markets with adverse selection. Economic Theory 7 (2), 207–235.
Galí, Jordi, Gertler, Mark, López-Salido, J. David, 2007. Markups, gaps, and the welfare costs of business fluctuations. The Review of Economics and Statistics 89 (1), 44–59.
Galor, Oded, Weil, David N., 1996. The gender gap, fertility, and growth. American Economic Review 86 (3), 374–387.
Garibaldi, Pietro, 1998. Job Flow Dynamics and Firing Restrictions. European Economic Review 42 (2), 245–275.
Gertler, Mark, Trigari, Antonella, 2009. Unemployment fluctuations with staggered Nash wage bargaining. Journal of Political Economy 117 (1), 38–86.
Goldin, Claudia, Katz, Lawrence F., 2002. The power of the pill: oral contraceptives and women's career and marriage decisions. Journal of Political Economy 110 (4), 730–770.
Gordon, Donald F., 1974. A neo-classical theory of Keynesian unemployment. Economic Inquiry 12 (4), 431.
Gordon, Robert J., 1990. The measurement of durable goods prices. University of Chicago Press.
Gottschalk, Peter, Moffitt, Robert, 1994. The growth of earnings instability in the US labor market. Brookings Papers on Economic Activity 25 (1994-2), 217–272.
Greenwood, Jeremy, Seshadri, Ananth, Yorukoglu, Mehmet, 2005. Engines of liberation. Review of Economic Studies 72 (1), 109–133.
Guerrieri, Veronica, Shimer, Robert, Wright, Randall, 2009. Adverse selection in competitive search equilibrium. NBER Working Paper 14915.
Haefke, Christian, Sonntag, Marcus, van Rens, Thijs, 2008. Wage rigidity and job creation. IZA Discussion Paper 3714.
Hagedorn, Marcus, Manovskii, Iourii, 2008. The cyclical behavior of equilibrium unemployment and vacancies revisited. American Economic Review 98 (4), 1692–1706.
Hall, Robert E., 1997. Macroeconomic fluctuations and the allocation of time. Journal of Labor Economics 15 (1), 223–250.
Hall, Robert E., 2005. Employment fluctuations with equilibrium wage stickiness. American Economic Review 95 (1), 50–65.
Hall, Robert E., 2009. Reconciling cyclical movements in the marginal value of time and the marginal product of labor. Journal of Political Economy 117 (2), 281–323.
Hall, Robert E., Krueger, Alan B., 2008. Wage formation between newly hired workers and employers: survey evidence. NBER Working Paper 14329.
Hall, Robert E., Milgrom, Paul R., 2008. The limited influence of unemployment on the wage bargain. American Economic Review 98 (4), 1653–1674.
Hansen, Gary D., 1985. Indivisible labor and the business cycle. Journal of Monetary Economics 16 (3), 309–327.
Hansen, Gary D., Imrohoroğlu, Ayşe, 1992. The role of unemployment insurance in an economy with liquidity constraints and moral hazard. Journal of Political Economy 100 (1), 118–142.
Hopenhayn, Hugo A., Nicolini, Juan Pablo, 1997. Optimal unemployment insurance. Journal of Political economy 105 (2), 412–438.
Hornstein, Andreas, Krusell, Per, Violante, Giovanni L., 2007. Technology-policy interaction in frictional labor-markets. Review of Economic Studies 74 (4), 1089–1124.
Hornstein, Andreas, Krusell, Per, Violante, Giovanni L., 2009. Frictional wage dispersion in search models: a quantitative assessment. Federal Reserve Bank of Richmond Manuscript.
Hosios, Arthur J., 1990. On the efficiency of matching and related models of search and unemployment. Review of Economic Studies 57 (2), 279–298.

Inderst, Roman, Wambach, Achim, 2001. Competitive insurance markets under adverse selection and capacity constraints. European Economic Review 45 (10), 1981–1992.

Inderst, Roman, Wambach, Achim, 2002. Capacity constrained firms in (labor) markets with adverse selection. Economic Theory 19 (3), 525–548.

Johnson, G.E., Layard, Richard, 1986. The natural rate of unemployment: explanation and policy. Handbook of Labor Economics 2, 921–999.

Jones, Larry E., Manuelli, Rodolfo E., McGrattan, Ellen R., 2003. Why are married women working so much? Federal Reserve Bank of Minneapolis Staff Report 317.

Katz, Lawrence F., Autor, David H., 1999. Changes in the wage structure and earnings inequality. In: Ashenfelter, Orley, Card, David E. (Eds.), Handbook of Labor Economics, vol. 3. Elsevier, pp. 1463–1555 (Chapter 26).

Kennan, John, 2010. Private information, wage bargaining and employment fluctuations. Review of Economic Studies 77 (2), 633–664.

Kitao, Sagiri, Ljungqvist, Lars Sargent, Thomas J., 2009. A life cycle model of trans-atlantic employment experiences. NYU mimeo.

Krugman, Paul, 1994. Past and prospective causes of high unemployment. Federal Reserve Bank of Kansas City Economic Review 79 (4), 23–43.

Krusell, Per, Mukoyama, Toshihiko, Rogerson, Richard, Şahin, Ayşegül, 2009. Aggregate labor market outcomes: the role of choice and chance. NBER Working Paper 15252.

Kugler, Adriana D., Saint-Paul, Gilles, 2004. How do firing costs affect worker flows in a world with adverse selection?. Journal of Labor Economics 22 (3), 553–584.

Kydland, Finn E., Prescott, Edward C., 1982. Time to build and aggregate fluctuations. Econometrica: Journal of the Econometric Society 50 (6), 1345–1370.

Layard, Richard, Nickell, Stephen, Jackman, Richard, 2005. Unemployment: Macroeconomic Performance and the Labour Market. Oxford University Press.

Lentz, Rasmus, Mortensen, Dale T., 2010. Labor market models of worker and firm heterogeneity. University of Wisconsin at Madison mimeo.

Lilien, David M., 1982. Sectoral shifts and cyclical unemployment. The Journal of Political Economy 90 (4), 777–793.

Lilien, David M., Hall, Robert E., 1986. Cyclical fluctuations in the labor market. In: Ashenfelter, Orley, Layard, Richard (Eds.), Handbook of Labor Economics, vol. 2. Elsevier, pp. 1001–1038 (Chapter 17).

Lippman, Steven A., McCall, John J., 1976. The economics of job search: a survey. Economic Inquiry 14 (2), 155–189.

Ljungqvist, Lars, 2002. How do lay-off costs affect employment?. Economic Journal 112 (482), 829–853.

Ljungqvist, Lars, Sargent, Thomas J., 1998. The European unemployment dilemma. Journal of Political Economy 106 (3), 514–550.

Ljungqvist, Lars, Sargent, Thomas J., 2004. European unemployment and turbulence revisited in a matching model. Journal of the European Economic Association 2 (2–3), 456–468.

Ljungqvist, Lars, Sargent, Thomas J., 2007. Understanding european unemployment with matching and search-island models. Journal of Monetary Economics 54 (8), 2139–2179.

Ljungqvist, Lars, Sargent, Thomas J., 2008. Two questions about European unemployment. Econometrica 76 (1), 1–29.

Lopes de Melo, Rafael, 2009. Sorting in the labor market: theory and measurement. University of Chicago mimeo.

Lucas Jr., Robert E., Prescott, Edward C., 1974. Equilibrium search and unemployment. Journal of Economic Theory 7 (2), 188–209.

MacLeod, W. Bentley, Malcomson, James M., 1989. Implicit contracts, incentive compatibility, and involuntary unemployment. Econometrica 57 (2), 447–480.

Manning, Alan, 2011. Imperfect competition in the labor market. In: New Developments and Research on Labor Markets. In: Ashenfelter, Orley, Card, David E. (Eds.), Handbook of Labor Economics, vol. 4b. pp. 973–1041.

Marimon, Ramon, Zilibotti, Fabrizio, 1999. Unemployment vs. mismatch of talents: reconsidering unemployment benefits. Economic Journal 109 (455), 266–291.

Mazumder, Bhashkar, 2007. New Evidence on Labor Market Dynamics over the Business Cycle. Federal Reserve Bank of Chicago Economic Perspectives. (1Q): 36–46.
McDaniel, Cara, 2009. Hours worked in the OECD 1960-2003: driving forces and propagation mechanisms. Kenyon College mimeo.
Menzio, Guido, Shi, Shouyong, 2009. Efficient search on the job and the business cycle. NBER Working Paper 14905.
Merz, Monika, 1995. Search in the labor market and the real business cycle. Journal of Monetary Economics 36 (2), 269–300.
Moen, Espen R., 1997. Competitive search equilibrium. Journal of Political Economy 105 (2), 385–411.
Montgomery, James D., 1991. Equilibrium wage dispersion and interindustry wage differentials. Quarterly Journal of Economics 106 (1), 163–179.
Mortensen, Dale T., 1982. Property rights and efficiency in mating, racing, and related games. American Economic Review 72 (5), 968–979.
Mortensen, Dale T., 1987. Job search and labor market analysis. In: Ashenfelter, Orley, Layard, Richard (Eds.), Handbook of Labor Economics, vol. 2. Elsevier, pp. 849–919 (Chapter 15).
Mortensen, Dale T., 1992. Search theory and macroeconomics: a review essay. Journal of Monetary Economics 29 (1), 163–167.
Mortensen, Dale T., 2003. How monopsonistic is the (Danish) labor market. In: Aghion, Philippe, Frydman, Roman, Stiglitz, Joseph E., Woodford, Michael (Eds.), Knowledge, Information and Expectations in Modern Macroeconomics. Princeton University Press, pp. 285–308 (Chapter 14).
Mortensen, Dale T., 2005. Wage Dispersion: Why are Similar Workers Paid Differently? The MIT Press.
Mortensen, Dale T., Nagypál, Éva, 2007. More on unemployment and vacancy fluctuations. Review of Economic Dynamics 10 (3), 327–347.
Mortensen, Dale T., Pissarides, Christopher A., 1994. Job creation and job destruction in the theory of unemployment. Review of Economic Studies 61 (3), 397–415.
Mortensen, Dale T., Pissarides, Christopher A., 1999a. Job reallocation, employment fluctuations and unemployment. In: Taylor, John B., Woodford, Michael (Eds.), Handbook of Macroeconomics, vol. 1. Elsevier, pp. 1171–1228 (Chapter 18).
Mortensen, Dale T., Pissarides, Christopher A., 1999b. New developments in models of search in the labor market. In: Ashenfelter, Orley, Card, David E. (Eds.), Handbook of Labor Economics, vol. 3. Elsevier, pp. 2567–2627 (Chapter 39).
Mortensen, Dale T., Pissarides, Christopher A., 1999c. Unemployment responses to 'skill-biased' technology shocks: the role of labour market policy. Economic Journal 109 (455), 242–265.
Mortensen, Dale T., Wright, Randall, 2002. Competitive pricing and efficiency in search equilibrium. International Economic Review 43 (1), 1–20.
Moscarini, Giuseppe, Postel-Vinay, Fabien, 2008. The timing of labor market expansions: new facts and a new hypothesis. NBER Macroeconomics Annual 23, 1–51.
Mulligan, Casey B., 2002. A century of labor-leisure distortions. NBER Working Paper 8774.
Nagypál, Éva, 2007. Labor-market fluctuations and on-the-job search. Northwestern University Manuscript.
Nickell, Stephen, Layard, Richard, 1999. Labor market institutions and economic performance. In: Ashenfelter, Orley, Card, David E. (Eds.), Handbook of Labor Economics, vol. 3. Elsevier, pp. 3029–3084 (Chapter 46).
Nickell, Stephen, Nunziata, Luca, Ochel, Wolfgang, 2005. Unemployment in the OECD since the 1960s. what do we know?. Economic Journal 115 (500), 1–27.
Ohanian, Lee, Raffo, Andrea, Rogerson, Richard, 2008. Long-term changes in labor supply and taxes: evidence from OECD countries, 1956–2004. Journal of Monetary Economics 55 (8), 1353–1362.
Olivetti, Claudia, 2006. Changes in women's hours of market work: the role of returns to experience. Review of Economic Dynamics 9 (4), 557–587.
Parkin, Michael, 1988. A method for determining whether parameters in aggregative models are structural. Carnegie-Rochester Conference Series on Public Policy 29, 215–252.
Peters, Michael, 1991. Ex ante price offers in matching games: non-Steady States. Econometrica 59 (5), 1425–1454.

Petrongolo, Barbara, Pissarides, Christopher A., 2001. Looking into the black box: a survey of the matching function. Journal of Economic Literature 39 (2), 390–431.
Petrongolo, Barbara, Pissarides, Christopher A., 2008. The ins and outs of European unemployment. American Economic Review 98 (2), 256–262.
Pissarides, Christopher A., 1984. Efficient job rejection. Economic Journal 94, 97–108.
Pissarides, Christopher A., 1985. Short-run equilibrium dynamics of unemployment, vacancies, and real wages. American Economic Review 75 (4), 676–690.
Pissarides, Christopher A., 2000. Equilibrium Unemployment Theory. MIT Press.
Pissarides, Christopher A., 2007. Unemployment and hours of work: the North Atlantic divide revisited. International Economic Review 48 (1), 1–36.
Pissarides, Christopher A., 2009. The unemployment volatility puzzle: Is wage stickiness the answer?. Econometrica 77 (5), 1339–1369.
Ponomareva, Natalia, Sheen, Jeffrey 2009. Cyclical flows in australian labour markets. Macquarie University mimeo.
Postel-Vinay, Fabien, Robin, Jean-Marc, 2002. Equilibrium wage dispersion with worker and employer heterogeneity. Econometrica 70 (6), 2295–2350.
Poterba, James M., Summers, Lawrence H., 1986. Reporting errors and labor market dynamics. Econometrica 54 (6), 1319–1338.
Prescott, Edward C., 2004. Why do Americans work so much more than Europeans? Federal Reserve Bank of Minneapolis Quarterly Review 28 (1), 2–13.
Pries, Michael, Rogerson, Richard, 2005. Hiring policies, labor market institutions, and labor market flows. Journal of Political Economy 113 (4), 811–839.
Ragan, Kelly S., 2004. Fiscal policy and the family: explaining labor supply in a model with household production. University of Chicago mimeo.
Ramey, Garey, Watson, Joel, 1997. Contractual fragility, job destruction, and business cycles. Quarterly Journal of Economics 112 (3), 873–911.
Ramey, Valerie A., Francis, Neville, 2009. A century of work and leisure. American Economic Journal: Macroeconomics 1 (2), 189–224.
Rogerson, Richard, 2007. Taxation and market work: Is Scandinavia an outlier? Economic Theory 32 (1), 59–85.
Rogerson, Richard, 2008. Structural transformation and the deterioration of European labor market outcomes. Journal of Political Economy 116 (2), 235–259.
Rogerson, Richard, Shimer, Robert, Wright, Randall, 2005. Search-theoretic models of the labor market: a survey. Journal of Economic Literature 43 (4), 959–988.
Rotemberg, Julio J., Woodford, Michael, 1991. Markups and the business cycle. NBER Macroeconomics Annual 6, 63–129.
Rotemberg, Julio J., Woodford, Michael, 1999. The cyclical behavior of prices and costs. In: Taylor, John B., Woodford, Michael (Eds.), Handbook of Macroeconomics. In: Handbook of Macroeconomics, vol. 1. Elsevier, pp. 1051–1135 (Chapter 16).
Rudanko, Leena, 2009. Labor market dynamics under long-term wage contracting. Journal of Monetary Economics 56 (2), 170–183.
Shapiro, Carl, Stiglitz, Joseph E., 1984. Equilibrium unemployment as a worker discipline device. American Economic Review 74 (3), 433–444.
Shavell, Shavell, Weiss, Laurence, 1979. The optimal payment of unemployment insurance benefits over time. Journal of Political Economy 87 (6), 1347–1362.
Shi, Shouyong, 2002. A directed search model of inequality with heterogeneous skills and skill-biased technology. Review of Economic Studies 69 (2), 467–491.
Shimer, Robert, 1996. Contracts in a frictional labor market. MIT mimeo.
Shimer, Robert, 2004. The consequences of rigid wages in search models. Journal of the European Economic Association 2 (2–3), 469–479.
Shimer, Robert, 2005a. The assignment of workers to jobs in an economy with coordination frictions. Journal of Political Economy 113 (5), 996–1025.
Shimer, Robert, 2005b. The cyclical behavior of equilibrium unemployment and vacancies. American Economic Review 95 (1), 25–49.

Shimer, Robert, 2007. Reassessing the Ins and Outs of Unemployment. mimeo.
Shimer, Robert, 2010. Labor Markets and Business Cycles. Princeton University Press.
Shimer, Robert, Smith, Lones, 2000. Assortative matching and search. Econometrica 68 (2), 343–369.
Shimer, Robert, Werning, Iván, 2008. Liquidity and insurance for the unemployed. American Economic Review 98 (5), 1922–1942.
Silva, José, Toledo, Manuel, 2009. Labor turnover costs and the cyclical behavior of vacancies and unemployment. Macroeconomic Dynamics 13 (Suppl. 1), 76–96.
Trabandt, Mathias, Uhlig, Harald 2009. How far are we from the slippery slope? the laffer curve revisited. NBER Working Paper 15343.
Trigari, Antonella, 2009. Equilibrium unemployment, job flows, and inflation dynamics. Journal of Money, Credit and Banking 41 (1), 1–33.
Veracierto, Marcelo, 2008. On the cyclical behavior of employment, unemployment and labor force participation. Journal of Monetary Economics 55 (6), 1143–1157.
Wolpin, Kenneth I., 1987. Estimating a structural search model: the transition from school to work. Econometrica 55 (4), 801–817.
Working, Holbrook, 1960. Note on the correlation of first differences of averages in a random chain. Econometrica 28 (4), 916–918.

CHAPTER 8

Extrinsic Rewards and Intrinsic Motives: Standard and Behavioral Approaches to Agency and Labor Markets☆

James B. Rebitzer [*], **Lowell J. Taylor** [**]

[*] Boston University and NBER

[**] Carnegie Mellon University

Contents

☆ We are grateful for very helpful comments from Linda Babcock.

Handbook of Labor Economics, Volume 4a
ISSN 0169-7218, DOI 10.1016/S0169-7218(11)04114-1

Abstract

Employers structure pay and employment relationships to mitigate agency problems. A large literature in economics documents how the resolution of these problems shapes personnel policies and labor markets. For the most part, the study of agency in employment relationships relies on highly stylized assumptions regarding human motivation, e.g., that employees seek to earn as much money as possible with minimal effort. In this essay, we explore the consequences of introducing behavioral complexity and realism into models of agency within organizations. Specifically, we assess the insights gained by allowing employees to be guided by such motivations as the desire to compare favorably to others, the aspiration to contribute to intrinsically worthwhile goals, and the inclination to reciprocate generosity or exact retribution for perceived wrongs. More provocatively, from the standpoint of standard economics, we also consider the possibility that people are driven, in ways that may be opaque even to themselves, by the desire to earn social esteem or to shape and reinforce identity.

JEL classification: J41; A12

Keywords: Principal agent models; Intrinsic and extrinsic motivation

1. INTRODUCTION

Many of the most widely-discussed and contentious issues facing the US economy concern the use of incentives to solve agency problems. Consider, for example, the problem of reforming the financial system following the recent collapse of financial markets. Explanations for the crash, as well as proposed strategies for effective reform, pivot around the adequacy of high powered financial incentives for ensuring that CEOs, rating agencies, financial advisors and brokers act in the interests of their constituents. Similarly, widely discussed proposals for improving health care quality and reducing costs involve "pay for performance" programs that reward the provision of "cost effective" health care. A growing literature in the economics of education is also exploring the efficacy of rewarding teachers for enhancing student performance.

In these debates, advocates argue that high powered incentives are necessary to get important work done efficiently. Thus, even the very large bonuses to top executives and elite financial engineers are "worth it" in the sense that expected gains from improved performance easily exceed the amount paid out. Critics counter that advocates for high powered incentive systems misunderstand human motivation. High powered incentives are unnecessary because appropriately motivated, selected and socialized agents will perform as well or better when stakes are lower. From this perspective incentives are inefficient because they generate unnecessary and potentially costly inequality within work groups or peer groups and because they needlessly divert agents' attention away from valuable aspects of their jobs that are hard to monitor and reward. In extreme cases, powerful incentives can cause agents to engage in malfeasance. Even more provocatively, some critics argue that the provision of extrinsic incentives undermine

agents' intrinsic motives and, in this way, worsen the incentive problem they are designed to solve.

Although advocates and critics may not be aware of it, the public controversies about incentive pay are essentially disputes about the appropriate specification of a workhorse economic model: the principal agent model. In its basic form, this model supports the idea that extrinsic rewards can be an efficient means of motivating agents. The claims of the critics are supported, however, when more realistic—and *ad hoc*—behavioral assumptions are introduced. Close examination of principal agent models reveals, furthermore, that debates about agency have implications far beyond issues of optimal incentive design. Indeed, the strategies firms adopt to resolve agency problems can have profound effects on labor markets broadly, affecting gender and racial inequality, labor market segmentation and unemployment.

In this chapter we review and analyze the principal agent model from a behavioral perspective. Although the literature is vast, our task is made simpler by the fact that conventional and behavioral principal agent models share a similar structure. In the simplest conventional models an agent is assumed to have utility that is increasing in earnings and decreasing in the provision of effort. Given this utility function, the principal can assess how the agent will react to a given reward structure, and can often link rewards to performance in a way that induces agents to supply efficient levels of effort—even if agents are entirely self-interested and even if measures of performance are noisy and imperfect. Behavioral models employ the same structure, but modify the agent's utility function to include additional psychological factors.

To complicate matters, in applications it is not sufficient to study an isolated agent responding to the policies of an isolated principal. Agents typically work as part of larger groups within organizations and society more broadly, and this can have important implications for the design of reward structures, especially when people have other-regarding preferences, care about inequality, or belong to groups with established norms of appropriate behavior. The policies adopted by firms may also have unexpected effects on labor market outcomes (e.g., can affect unemployment rates) and these outcomes may, in turn, alter the optimal policies of individual firms.

A second complication for conventional principal agent models is that pay structures often perform "double duty," e.g., they must resolve both a motivation problem and some other problem. For instance, principals (firms) might adopt compensation and employment practices that signal the principal's ability to make good on promises to agents. Conversely, employment practices might be designed so as to allow agents to signal some hidden characteristic about themselves, as in "rat race" models in which individuals provide "excessive" work hours as a means of signalling an otherwise unobserved personal inclination to work hard. Pay structure also performs "double duty" when workers must attend to multiple tasks. In these situations, rewards for high performance along one dimension draw effort and attention away from other valuable

dimensions of performance. In turn, principals must be careful in the assignment of multiple tasks, and might also want to tilt toward lower-powered incentives.

Just as in conventional principal agent theory, "double duty" incentives play an important role in behavioral principal agent models. For instance, in behavioral models pay structures not only elicit effort, but also influence employee perceptions of the legitimacy of the reward structure. Indeed, from a behavioral perspective, a key task of management is to persuade employees of the legitimacy of tasks and rewards and so to help socialize them into the mutually reinforcing expectations of the group.

Personnel policies also do "double duty" if, as is commonly assumed in behavioral models, agents have intrinsic motivation. For example, when agents differ in their intrinsic alignment with an organization's mission, firms with especially evocative missions may design their pay structures so as to attract workers who identify with that mission. Signaling variants of "double duty" incentives are also prevalent in behavioral models of motivation. For example, firms might use compensation policies to signal workers about the likely motivations of co-workers, which can matter for workers who are inclined to conform to workplace norms. Signalling might be used also if the firm has hidden knowledge about a worker's suitability for a task. More provocatively, principals who are sensitive to psychological motivations might set up compensation policies to exploit the possibility that agents send signals to themselves, as a means of nurturing a sense of identity.

Our discussion proceeds as follows. In Section 2 we present a standard principal agent model. We begin with the simplest case—a single isolated agent working for an isolated principal. We then consider the complications that arise when we place this relationship into the context of a firm or a labor market. As we build our model in this section and throughout the paper, we refer to relevant empirical applications from experiments and from field data.

In Section 3 we introduce the problem of extrinsic rewards with "double duty" incentives. We discuss three applications: wages as a signal of firm fitness, rat races, and multi-tasking. In each case, the presence of double duty incentives greatly alters the market outcomes and employment relationships.

In Section 4 we introduce behavioral features to our agency model. To keep the discussion manageable, we focus on four issues: inequality aversion, the desire to reciprocate, behavioral norms, and identity/self-image. Many of the interesting applications in this section focus on professional settings and touch on professional norms.

Section 5 considers behavioral issues in the context of double duty incentives. The most interesting question we approach here is whether extrinsic rewards might "crowd out" valuable intrinsic motivation.

We conclude the paper by highlighting what we see as promising areas for future research.

2. AGENCY AND EXTRINSIC REWARDS

2.1. A simple agency problem

There are a great many interactions in the labor market that can be fruitfully examined as a principal agent problem—an interaction in which the principal uses a reward structure to motivate an agent to pursue some desired action. As a baseline example, consider a principal who seeks to maximize profit, which depends on the "effort" of an agent and the compensation given to that agent, as specified by

$$g(e) - w. \tag{1}$$

In (1), $g(e)$ is the value produced for the principal as a consequence of an agent's effort e (which can be represented as a non-negative scalar) and w is compensation given to the agent. We assume that $g'(e) > 0$ and $g''(e) < 0$ exist and are continuous. As for the agent, we assume simply that utility is $w - e$. Thus e is the money metric disutility of taking the action that benefits the principal. What makes this problem interesting is that the principal cannot directly observe effort.

Although information asymmetry is essential to our story, to set basic ideas, we ask initially what the solution would be if the principal could observe the agent's effort, and write a contract specifying effort and wage. The firm would then simply maximize (1), subject to the agent's *participation constraint*, which specifies that the utility resulting from the agreed-upon wage and effort must equal or exceed the utility available to the agent elsewhere, i.e., $w - e \geq v$. The principal finds it most profitable to operate with this latter constraint binding, so it immediately follows that the solution to this constrained maximization problem entails

$$g'(e^*) = 1. \tag{2}$$

This outcome is efficient: the marginal value of additional effort equals the marginal cost to the worker of supplying the effort. This agreed-upon effort level is the same as if the agent worked for herself. The resulting wage is $w^* = v + e^*$.

When, instead, the principal does *not* observe the agent's effort level, the principal must find an incentive mechanism to induce the desired effort level. One possibility is that pay can be conditioned on the value of output $g(e)$. In some instances, though, neither output nor effort are readily observable. We consider such a case. We suppose, instead, that the principal has a noisy signal of effort,

$$x = e + \epsilon, \tag{3}$$

where ϵ is drawn from a differentiable, symmetric, single-peaked density f (with corresponding cumulative density F). There are variety of possible interpretations consistent with this set-up. For example, x might be some objective measure of

performance, and ϵ is simply luck. Alternatively, we might interpret x as a principal's impression or opinion of how well a worker is performing, so x is unobservable by outside parties. The ϵ term in this latter case captures miscommunication and misinterpretation of effort.

We start with the case in which the realized value of x is common knowledge, adopting the assumption that the principal can be trusted to honor commitments in which compensation w is conditioned on x. In this case, there are many incentive schemes that will do. To set the stage for results to come, we work through a particularly simple scheme: we assume that the principal commits to a policy of paying wage w_0 if the observed performance x falls below some cut-off $\bar{x}$, and paying $w_1 > w_0$ if x is above that cut-off.[1]

To be clear, we have the following timing in mind: (i) The principal announces the policy (including $\bar{x}$), and posts w_0 and w_1. (ii) The agent decides whether to accept the job, and if she does, takes hidden action e. (iii) Nature plays x. (iv) Given x, the firm pays the agreed-upon wage.

We can easily solve for the optimal wage policy, (w_0^*, w_1^*). Conceptually, the first step is to account for the agent's *best response* to the wage policy. At effort level e, the probability of earning w_0 is $F(\bar{x} - e)$ and the probability of earning w_1 is $1 - F(\bar{x} - e)$. So the agent wants to maximize

$$w_0 F(\bar{x} - e) + w_1 [1 - F(\bar{x} - e)] - e, \tag{4}$$

which leads to the best response, $\hat{e}$, that solves

$$(w_1 - w_0) f(\bar{x} - \hat{e}) - 1 = 0. \tag{5}$$

From this last expression, we notice that the best response is a function of the difference between the higher wage and the lower wage, say $b \equiv w_1 - w_0$ (b is the "bonus" that accompanies the high-performance outcome).[2] Thus we can write $\hat{e}(b)$, noting, for future reference that

$$\hat{e}'(b) = f(\bar{x} - \hat{e})/[b f'(\bar{x} - \hat{e})] > 0 \tag{6}$$

(under the assumption that the second order condition holds). This makes sense; higher-powered incentives elicit greater effort.

[1] For the moment we take the cut-off as given, but as will become apparent, in this particular model it is important that the chosen cut-off be lower than the hoped-for level of performance, e^*.

[2] The second order condition is $-bf'(\bar{x} - \hat{e}) < 0$, which is satisfied if $\hat{e}$ exceeds the cut-off $\bar{x}$. Hence our observation in the previous footnote that $\bar{x}$ must be below the hoped-for level of effort.

Next, the principal must account for a *participation constraint*. The agent accepts the job only if the expected wage equals or exceeds the agent's opportunity cost:

$$w_0 F(\bar{x} - e) + w_1 [1 - F(\bar{x} - e)] \geq v + \hat{e}(b). \tag{7}$$

The principal's problem then turns out to be straightforward. Expected profit is output minus the expected wage, and given that the participation constraint binds, this is just

$$g(\hat{e}(b)) - [v + \hat{e}(b)]. \tag{8}$$

The first order condition to the principal's profit maximization problem is

$$[g'(\hat{e}(b^*)) - 1]\hat{e}'(b^*) = 0. \tag{9}$$

Above, we noted that $\hat{e}'(b) > 0$ for any best response, so the elicited effort level, $e^* \equiv \hat{e}(b^*)$, described by (9) solves

$$g'(e^*) = 1. \tag{10}$$

The solution thus entails the efficient level of effort, as in (2). The principal pins down b^* using (5), which can be read as giving b^* as an implicit function of e^*. Finally, given b^* and e^*, the firm sets the base wage w_0^* to be as low as possible, while still meeting the participation constraint (7).

We have obviously chosen to work out an extremely simple case as our prototypical principal-agent model.[3] As simple as the model is, it is nevertheless sufficient to make the point that a solution to the agency problem entails a strategy of conditioning pay on observed productivity. This reward structure can elicit efficient effort levels even when agents are entirely self-interested and when performance measures are noisy and imperfect.

2.2. Agency matters

In this section we demonstrate the value of thinking carefully about agency in the context of three labor market applications: (i) CEO compensation, a case in which there is a single agent, (ii) personnel policies in a firm, which involves a single principal seeking to motivate multiple agents, and (iii) unemployment and labor market segmentation that can arise in labor markets in which multiple principals compete for agents, and in which the motivation problem is addressed by the threat of dismissal. In each of these three cases, solutions to the principal agent problem are seen to have important social consequences.

[3] Natural alternative conceptions would allow for risk aversion, as in the Holmström's (1979) classic paper.

In each case also, empirical evidence indicates that anomalies exist that point to the importance of behavioral aspects that are not included in the standard principal agent set-up.

2.2.1. CEO compensation

In advanced economies with modern financial systems, top executives of publicly traded corporations and large financial firms play a central role in the allocation of productive resources. Thus the reward structure under which these executives operate is of considerable economic interest. The rapid increase in the pay of CEOs since the early 1980s is also one of the most striking labor market developments of the past 25 years. These pay increases have contributed in an important way to growing income inequality (Levy and Temin, 2007) and they have also been the target of intense public controversy. The rise in CEO compensation is inextricably linked to agency issues because most of the changes in pay are the result of increasing grants of stocks and stock options. For example, Hall and Liebman (1998) report the median elasticity of CEO compensation with respect to firm stock market performance more than tripled between 1980 and 1994, largely because of the rapid rise in stock based compensation. Bebchuk and Grinstein (2005) document a continuing rapid growth in equity-based compensation for CEOs and top five executives through 2003.

One of the great appeals of the principal agent model is that it tells us what efficient CEO reward structures ought to look like. A central prediction of the model is that the efficient reward structure for CEOs and other top executives should have higher levels of expected pay and higher incentive intensity than for other employees. As a simple formal example, suppose that the value to a firm of a particular agent's effort is $g(e) = \theta h(e)$, where $h(e)$ is a concave function increasing in e, and θ is a positive constant that differs across individuals within an organization, depending on the importance of that individual's job to the organization's profitability. CEOs and top executives will typically have the highest values of θ. At the efficient level of effort, $\theta h'(e^*) = 1$ and $\theta h''(e^*) < 0$ (assuming that the second order condition holds), so $\frac{\mathrm{d}e^*}{\mathrm{d}\theta} = \frac{-h'(e^*)}{\theta h''(e^*)} > 0$. Thus effort expectations are highest for CEOs and top executives and, because the agent's "best response" effort is increasing in the size of the bonus, b will also be highest for them. The size of the bonus, b, is likely to be very large, particularly in an environment in which it is difficult to assess the CEO's absolute performance.

One way of expressing the agency model presented above is that compensation should be set so that any agent becomes (at the margin and in expected value) the *residual claimant* with respect to her contributions to the firm; her own personal fortune rises or falls as a consequence of the value of the actions she undertakes on behalf of her firm, $g(e)$. Now in our set-up above, the firm conditions compensation on an imperfect measure of e, under an assumption about the infeasibility of measuring $g(e)$ itself for a typical employee. But in the case of the CEO, her actions might be so consequential to the firm

that her contribution essentially represents *firm profit* itself. If so, perhaps the ideal contract would make her, roughly, the residual claimant to the entire corporation. To make that happen, one would want to tie CEO compensation tightly to firm profitability (i.e., stock values) and then give the CEO unlimited latitude with regard to actions she takes on behalf of the firm.

At first blush, incentives for CEOs appear to match well the predictions of the bare bones principal agent model. Top executives earn multi-million dollar salaries and the bulk of their compensation comes in the form of pay linked to stock-based performance measures, as one would expect if stock markets are efficient evaluators of firm value. Indeed, empirical analysis by Gayle and Miller (2009) indicates that the pattern of rising CEO pay and the rising incentive intensity of this pay over a sixty year period can be explained largely by parameters emphasized in the principal agent model: increasing losses to the firm from CEOs pursuing their own goals rather than value maximization, and the deteriorating value of stock performance as a signal of CEO effort. The former is the result largely of the increasing size of firms.

While it is clear that CEOs ought to function under higher-powered incentives than other employees, it is not clear if compensation boards are setting incentives properly. In their seminal article, "Performance Pay and Top-Management Incentives," Jensen and Murphy (1990) estimate that during the period 1969 through 1983, CEO wealth increased by only \$3.25 in response to a \$1000 increase in firm value. This number would seem to offer a *prima facie* case for CEOs having inadequate incentives to increase shareholder value. Hall and Liebman (1998) present empirical evidence that in fact there is a substantially tighter link between CEO compensation and firm value, particularly when they examine more recent periods (1980-1994).[4]

Still, in large corporations CEOs are far from being residual claimants. As Hall and Liebman (1998) suggest, this might pose little problem for the proper alignment of some CEO actions but create large problems for others. For example, a CEO who receives \$1 in compensation per \$100 value created for a firm might be sufficiently motivated to make smart, carefully-reasoned decisions about resource allocation to proposed projects. But this same CEO gets an effective 99% discount on a multi-million jet purchased by the firm for his own use.

This latter point is easily illustrated with a slight modification to our baseline principal agent model. Suppose, now, that firm profit is

$$g(e) - w_S - w_N, \tag{11}$$

where now w_S is the CEO's salary and w_N is the non-salary cost that results from

[4] In particular, they estimate that during the period 1980-1994 a typical CEO whose actions caused the firm to move from the 30th percentile of annual returns to the 70th percentile enjoyed an increase in annual compensation of 1 to 4 million dollars (1994 dollars), mostly through the increased value of the CEO's stocks and stock options. For stellar performance the increase in CEO wealth was estimated to be much higher.

the CEO's actions that increase the CEO's welfare at the expense of the firm (e.g., expenditures on a jet for CEO use). We now let the CEO's utility be $w_S + u(w_N) - e$, where $u(w_N)$ is the money metric value to the CEO of non-salary expenditures—a function that is obviously increasing in w_N. We also expect $u'(0) > 1$ and $u''(w_N) < 0$.[5] If the firm could observe and direct e and w_N, it would choose e^* and w_N^* so that

$$g'(e^*) = 1 \quad \text{and} \quad u'(w_N^*) = 1. \tag{12}$$

Suppose instead the firm sets the variable component of the CEO's compensation equal to the share s of the firm's profit, i.e., $s[g(e) - w_N]$. The best response here will entail the CEO choosing

$$g'(e^{**}) = 1/s \quad \text{and} \quad u'(w_N^{**}) = s. \tag{13}$$

Comparison of (12) and (13) demonstrates the problem: If $s < 1$, we have too little CEO effort on behalf of the firm and too much squandering of resources on the CEO.

How should the corporation's compensation board respond? One argument is that s must be driven ever closer to 1, even if this entails a substantial direct surplus transfer to the CEO. An alternative might entail the judicious combination of monitoring and more-narrowly directed incentives—a process that would likely play on the hope that hard-to-observe excessive levels of w_N by the CEO would be limited by shame or a sense of obligation to shareholders. This latter strategy can only be studied in a set-up that takes such behavioral aspects into account.

In any event, it is infeasible for firms to set up pay structures in which CEOs literally become residual claimants. The issue at hand is readily visible in our baseline principal agent model. As we note above, if it is optimal to induce a CEO to exert a very high level of effort on behalf of the firm, it is necessary also to have a very high bonus. With a binding participation constraint, this means that the contract will specify *negative* base pay. If this is infeasible, i.e., if the CEO cannot be compelled to pay the firm when performance is poor, it will be necessary to modify the contract to take account of the CEO's *limited liability*.

A simple solution to compensation with limited liability might look like this: each year the CEO is offered high pay and is rewarded further by having her contract renewed for the following year if she has a high observed performance level, but she is dismissed if observed performance is insufficiently high. We characterize this solution in detail in Section 2.2.3 below. Two key results that emerge in that analysis are germane here. First, the solution entails a surplus transfer from the firm to the CEO; the necessity of having a high-powered incentive leads to an especially high salary for the CEO when there is

[5] Concavity is natural here. The assumption $u'(0) > 1$ simply allows for the fact that *some* expenditures on CEO wellbeing are efficient.

limited liability. Second, the more precise the principal is in assessing performance, the lower will be this surplus transfer.[6]

Our baseline principal agent model therefore predicts that profit-maximizing principals will, whenever possible, seek to reward performance and not luck.[7] There is persuasive evidence that, to the contrary, at least some CEOs are rewarded for observable luck. In particular, Bertrand and Mullainathan (2001) show that because of the way many firms tie CEO compensation to stock market performance, "CEO pay in fact responds as much to a lucky dollar as to a general dollar." For example, increases in the world price of oil causes stock price increases in the oil industry. In the baseline principal agent model, such "luck" should have no impact on CEO pay, but in reality some CEOs are observed to reap handsome rewards simply because of such luck.

The use of high-intensity incentives through the use of stock options—a common way of tying compensation and firm outcomes—can create additional problems. If options are very far under water, their value as incentives degrades to near zero—obviously an undesirable state of affairs.[8] Conversely, when stock prices are just below the stock price, the payoff to even small increases in the stock price are huge, and this can create irresistible temptations to game the compensation system. Heron and Lie (2009), for example, estimate that 13.6% of all option grants to top executives during the period 1996-2005 were backdated or otherwise manipulated.

It is especially surprising that the use of stock options is a part of compensation even for many corporate employees below the top executive level. Hall and Murphy (2003) report that in S&P 500 corporations, roughly 90% of the outstanding options are granted to employees below the top five executives. This pattern is very hard to reconcile with principal agent models because the efforts of individual employees below the top five executives, it would seem, can have little direct influence on the price at which their company's stock trades. In this case, a stock-based compensation of any sort is likely to have little effect on effort levels. The use of stock-based compensation for lower-level employees is even harder to understand if agents are risk averse. Here again, the simple principal agent model appears to be inadequate for explaining compensation practices, perhaps because of the omission of important behavioral aspects.

2.2.2. Personnel policies

There have been a great many applications of the principal-agent model for the purpose of understanding compensation policies within firms more broadly. As a simple example,

[6] We prove this latter result formally in Section 2.2.3 for the important case in which there is an ongoing (multiple period) relationship between the principal (firm) and agent (CEO).

[7] Notice that this result holds even though we assume that agents are risk neutral. The result is reinforced, of course, if the agent is risk averse.

[8] Hall and Knox (2004) estimate that at the height of the bull market in 1999, roughly a third of executive options were under water. Companies often respond to this non-linearity in stock option returns by increasing option grants following stock price declines.

consider a profit-maximizing employer whose n workers produce output according to a production function, $Y = G(e^1, e^2, \ldots, e^n)$ per period, where e^i is worker i's effort. As above, output is increasing in effort: $\partial G/\partial e^i > 0$.

We continue to assume also that in a given period a worker chooses e and receives utility $w - e$, and that the firm cannot condition compensation on Y (or, in any event chooses not to). Importantly, for this application, we also assume that workers do not base their effort decisions on the effort or compensation of other workers.[9] Then we can treat the firm's agency problem with a given worker in terms of the function $g(e^i)$, which is the value of product that results from production $G(e^i, e^{(-i)})$, where $e^{(-i)}$ denotes effort levels of workers other than i. We assume that $g'(e^i)$ and $g''(e^i)$ are continuous, and that $g''(e^i) < 0$.

In setting up our baseline agency model in Section 2.1 we ignored an issue that is generally germane in the workplace: The indicator of performance, x in our model, is typically observed only by the manager and by workers within a firm, and thus cannot readily be used as the basis for forming contracts that can be enforced, say, by an outside court. Workers who understand that they will have no recourse if a manager violates the implicit agreement—"pay a bonus for high observed effort"—will logically decline to accept the proposed agreement. How might a firm proceed in this case? One possibility is to set up a competition among its n workers. Suppose, for example, that the firm cannot directly condition pay on x, but *can* commit to an evaluation process at the end of the period in which (i) workers are ordered on the basis of observed performance, and then (ii) the fraction P who are lowest-performing are paid w_0, while the remaining high-performing workers are paid $w_1 > w_0$. The key idea here is that while individual performance is not well observed, everyone can observe the agreed-upon reward structure and see if the firm is meeting that obligation.

We can easily find a Bayesian equilibrium in which all workers supply the same level of effort in response to the competition. Suppose that worker i believes that all other workers are going to play $\tilde{e}^{(-i)}$. Now what is her best response? The worker first uses her knowledge of P to accurately assess the cut-off value of observed performance, say $\tilde{x}$, which separates low- and high-performance assessments. That is, she takes note of the value $\tilde{x}$ that solves $F(\tilde{x} - \tilde{e}^{(-i)}) = P$. Given that value $\tilde{x}$, her optimal choice is to set effort level e^i so as to maximize

$$w_0 F(\tilde{x} - e^i) + w_1 \left[1 - F(\tilde{x} - e^i)\right] - e, \tag{14}$$

which leads to a best response given by

$$(w_1 - w_0) f(\tilde{x} - \hat{e}^i) - 1 = 0. \tag{15}$$

[9] This last assumption follows naturally enough, given the utility function we have specified. In our discussion of behavioral models in Section 4, we allow for the possibility that workers do care directly about the effort or compensation of other workers.

But this is exactly the worker best response we solved in our baseline example (compare (5) and (15)). Given this insight, it is easy to verify that the firm has a workable plan here: The firm starts by setting the "tournament prizes," (w_0, w_1), to be (w_0^*, w_1^*), as derived in our baseline example in Section 2.1. Then it chooses the fraction P^* so that $P^* w_0^* + (1 - P^*) w_1^*$ just satisfies the participation constraint (5). If worker i believes other workers are choosing effort level $\tilde{e}^{(-i)} = e^*$, she responds by also choosing e^*. All workers behave the same in equilibrium.[10]

The logic outlined in the preceding paragraph is the starting point of Malcomson's (1984) well-known paper on hierarchy and internal labor markets. He suggests that the "tournament prize" idea can be fruitful for thinking about the internal organization of the workplace. He works with a two-period model. In the first period of one's career, within a firm, each worker receives the same wage.[11] Then in the second period, the fraction $(1 - P)$ of workers who have been most successful as junior employees are promoted to high-paying jobs, while the fraction P who have been less successful are retained in low-paying jobs (at a wage that is high enough to keep them from moving to other firms). The tournament provides an extrinsic reward designed to elicit optimal effort from young workers.[12]

As Malcomson (1984) notes, the simple tournament model we have just outlined is consistent with some commonly observed features of organizations, e.g., that wage structures in organizations are often "hierarchical," with workers falling into distinct pay grades, that often workers in high-paid positions are promoted from within, that wages typically rise with seniority (perhaps by more than productivity), and that the variance of wages increases with seniority. Indeed, one of the major contributions of agency theory to labor economics is its ability to help explain the origin of firm wage policies and hence clarify the contribution that personnel practices make to shaping the wage structure.[13]

As was true in its application to CEO compensation, the first-order predictions of the agency model receive considerable empirical support, but there are anomalies that suggest the model may not offer an altogether satisfactory guide to understanding the internal structure of organizations.

[10] It is important here that the manager actually follows through on the promise to award the higher wage to workers who have the highest realized values of x. This might be sensible, especially if realized values of x are reasonably well known by people within the firm. After all, why wouldn't the manager want to reward to workers who have the highest performance outcomes? Having said that, if there is "favoritism" based on other criteria, the proposed incentive plan falls apart. For more on favoritism see Prendergast and Topel (1996).

[11] This wage solves a two-period participation constraint. The first period wage is low, possibly negative.

[12] In Malcomson's (1984) set-up, the agency problem is left unresolved for older workers. The point is that young workers can be motivated by the promise of future prizes (promotions, raises, etc.). Such incentives are less likely to be effective for workers nearing retirement. This is quite typical in agency-based models of "internal labor markets," and it doesn't substantially alter the basic insights generated in these models.

[13] There are now a number of insightful overviews of the topic, including Lazear (1998), Prendergast (1999), Malcomson (1999), Gibbons (1998), and Oyer and Shaefer's chapter in this *Handbook*.

To begin, it is important to recognize that extrinsic incentives do matter within organizations, often in exactly the way predicted by simple models of agency: Lazear (2000) found an increase in effort when a glass installer went from fixed pay to pay-for-performance. Kahn and Sherer (1990) document the effectiveness of an evaluation-and-bonus program at a manufacturing firm. Jacob (2005) shows that high-stakes testing in the Chicago Public Schools does alter teacher behavior—intensifying effort in improving student's test-specific skills, while substituting away from low-stakes subjects like science and social studies. Important work by Theodore Groves and John McMillan and their co-authors shows that strengthened incentives led to substantial productivity increases in Chinese industry and agriculture.[14] And, of course, there are many other examples in the literature.

Principal agent models also require that firms are choosing pay policies in an optimal way. It is hard to find direct evidence that pay policies are chosen in this way. Indeed, much of the literature showing that "incentives work" does so by exploiting the measured consequences of poorly designed incentives. That is, they clearly demonstrate that organizations—at least in some cases—do *not* choose incentives optimally. This is clear, for example, in Oyer (1998), which calls attention to the fact that salespeople seem to intensify effort at the end of the fiscal year if by doing so they can surpass performance thresholds and earn a bonus. At the organizational level, Courty and Marschke (2004) similarly demonstrate that a large government organization strategically reported performance outcomes to increase earned rewards, and did so at the expense of productive activities. In work with Martin Gaynor (Gaynor et al., 2004), we document the effects of an HMO's incentive contract designed to limit expenditures by physicians, but our identification strategy relies on the observation that a key feature of the incentive contract was implemented haphazardly. An even more extreme example is Jacob and Levitt's (2003) demonstration that public school teachers responded to a shift toward higher-powered incentives by cheating, e.g., by altering questions on standardized tests taken by students.

Some of the ancillary predictions of principal agent models also lack empirical support. In the two period model we present above, compensation in period 1 ought to move inversely to expected compensation in period 2—a result that follows directly from a two-period participation constraint. In an earlier paper studying law firms (Rebitzer and Taylor, 1995a) we tested this hypothesis. We find, contrary to the predictions of our principal agent model, that large law firms with extremely high second period compensation (in the form of the high income of partnership) also pay their starting associates high salaries relative to smaller firms. This would seem to indicate that successful law firms use some form of "rent sharing"—a strategy that emerges when we add such behavioral features as "inequality aversion" (in Section 4.1).

[14] See, e.g., McMillan et al. (1989), Groves et al. (1994) and Groves et al. (1995).

A particularly jarring feature of the minimalist agency model of personnel practices is the "irrelevance of *ex post* inequality." The compensation structure emerging from our model might indeed be termed "pay for luck" rather than "pay for performance." The principal and agent(s) know that the equilibrium effort level is e^*. Even so, it is important that pay be based on the measure of observed performance so as to provide the crucial extrinsic incentives. This feature—that rewards or punishments are based on an observed outcome, not on the actual behavior (even though those behaviors can be deduced)—is very common in game-theoretic approaches, including much of the work presented below. Anybody who has spoken with managers (or chaired an academic department) knows that people don't respond well when they are paid less than co-workers for what appears to be arbitrary, capricious or random reasons. This observation has been widely examined in the behavioral literature on agency, and we will discuss its implications in Section 4.

2.2.3. Involuntary unemployment and market segmentation

In 1984, Shapiro and Stiglitz set out an influential "efficiency wage" model that illustrates an important feature of agency models: the actions an individual firm takes to resolve an agency problem can give rise to important social costs when adopted throughout the market. In the case of efficiency wage models, the social costs are those arising from involuntary unemployment and labor market segmentation.[15] The set-up we present here is a recasting of the Shapiro and Stiglitz model taken from Ritter and Taylor (forthcoming).

We consider a market in which there are a large number of identical profit-maximizing employers, each of which faces the agency problem we outlined above. Each firm in the model is assumed to behave as outlined in Section 2.1: the idea is to pay well for "good outcomes" while penalizing workers for "bad outcomes" to the maximum extent possible. *Limited liability* is invoked through the assumption that the only penalty that the firm can implement is to dismiss a poorly performing worker. The motivation problem is resolved by employers making jobs sufficiently valuable that workers will provide effort so as to prevent dismissal.

To capture the idea that jobs have value, it is necessary to set the model up in a multi-period framework. The agency model outlined in Section 2.1 is thus assumed to pertain for each period indefinitely and workers are assumed to live indefinitely with a discount rate ρ.[16]

15 Similar points were also made in the important work of Bowles (1985). For additional theoretical development, see MacLeod and Malcomson (1989).

16 Valuable long-term employment relationships are central to these models, and thus so are shocks to employment. The model here can be enriched to allow for these exogenous shocks to employment relationships, but for simplicity we omit this feature.

The basic set-up

Employees are paid w for one unit of labor per period. In each period a worker chooses e, and this produces utility $w - e$. In this model, the alternative to employment is unemployment, which results in utility $v = 0$ in the period. The present value of being unemployed is V^u (which is not 0 because there is some prospect of being hired in the future). Hiring and termination are costless to the firm.

The model is a game between the firm and a worker with the following order of play in each round: (1) The firm offers a wage w. (2) The worker chooses effort level e. (3) Nature plays x using the distribution $f(\ )$. (4) The firm pays w. (5) The firm decides whether to retain the worker or end the game. We focus on the perfect Bayesian equilibrium in which the worker is retained if and only if x exceeds an endogenous threshold $\bar{x}$. We assume that $\bar{x}$ is common knowledge. (Workers can infer $\bar{x}$ by observing the frequency of terminations.)

The solution method mimics the steps we took in the simpler model above. In particular, we first find the worker's best response. Then we see how the firm will chose its personnel policies ($\bar{x}$ and w) in light of the worker's best response.

The worker's best response

Let $\hat{e}(w, \bar{x})$ be the worker's best response. To find that best response notice that for a person who chooses e in the current period, and then reverts to $\hat{e}$ in all future periods, lifetime utility is given by

$$V(e) = w - e + \frac{[F(\bar{x} - e)V^u + (1 - F(\bar{x} - e))V(\hat{e})]}{1 + \rho}. \tag{16}$$

The employee maximizes $V(e)$ by choosing $\hat{e} > 0$. For an interior solution, the first order condition is

$$\frac{[V(\hat{e}) - V^u]}{1 + \rho} f(\bar{x} - \hat{e}) - 1 = 0. \tag{17}$$

As in our baseline agency model, the second order condition holds when $f'(\bar{x} - e) > 0$. As we have noted, this incentive elicits effort because the job is valuable: $[V(\hat{e}) - V^u] > 0$. As is typical of models that invoke limited liability, the participation constraint does not bind.

Evaluating Eq. (16) at $e = \hat{e}$ and solving for $V(\hat{e})$, then substituting into the first order condition (17) produces

$$w = \hat{e} + \frac{\rho V^u}{1 + \rho} + \frac{\rho + F(\bar{x} - \hat{e})}{f(\bar{x} - \hat{e})}. \tag{18}$$

This last expression implicitly defines the worker's best response, $\hat{e}(w, \bar{x})$.

Firm profit maximization

Now we can turn to the firm's objective. It seeks to maximize profits, taking into account the worker's best response, i.e., maximizes

$$g(\hat{e}(w, \bar{x})) - w.$$

The solution can readily be found using the same steps we followed in solving the agency problem in Section 2.1. In this instance the optimal employment policy again induces the socially optimal performance level regardless of $f(\,)$:[17]

$$g'(e^*) = 1.$$

The noise in the environment does, however, affect the distribution of surplus. In particular, Ritter and Taylor (forthcoming) establish the following results: First, when the firm optimally chooses w and $\bar{x}$, the resulting probability of retention, $F(\bar{x}^* - \hat{e}^*)$, is unaffected by the level of σ^2 (the variance of the density $f(\,)$).[18] Second, the optimal wage does depend on σ^2, as follows:

$$w^* = e^* + \frac{\rho V^u}{1+\rho} + \frac{\phi(z^*)}{\phi'(z^*)}\sigma, \tag{19}$$

where z^* is a "standardized" random variable, $z^* = (\bar{x}^* - e^*)/\sigma$, and $\phi(\,)$ is the "standardized p.d.f.," i.e., the p.d.f. of $\frac{\epsilon}{\sigma}$. The more intractable the agency problem—the greater the value of σ—the higher is the wage required to achieve efficient effort and so the greater the surplus accruing to the worker.

Equilibrium unemployment

The equilibrium of the model we have just outlined generates unemployment.

Let V^* be the present value of lifetime utility of an employed individual who works at the optimal effort level. Recall that V^u is the expected lifetime utility for an individual who is unemployed. This utility level depends, clearly, on the probability of job acquisition. Let that rate be a.[19] Given that current-period utility of an unemployment person is zero, the expected lifetime utility of an unemployed individual is

$$V^u = 0 + \frac{[aV^* + (1-a)V^u]}{1+\rho}, \tag{20}$$

[17] This shows that efficiency wage motivation can lead to socially optimal effort levels, but this need not always be the case. For example, Allgulin and Ellingsen (2002) show that there can be distortions away from the socially optimal effort level when the principal has discretion over investments in monitoring.

[18] In fact, the probability of retention is shown to be $F^* = \frac{\phi(z^*)^2}{\phi'(z^*)} - \rho$.

[19] Note that the rate a is a known constant to any individual, but of course is endogenous to the economy as a whole. We solve for the equilibrium rate a shortly.

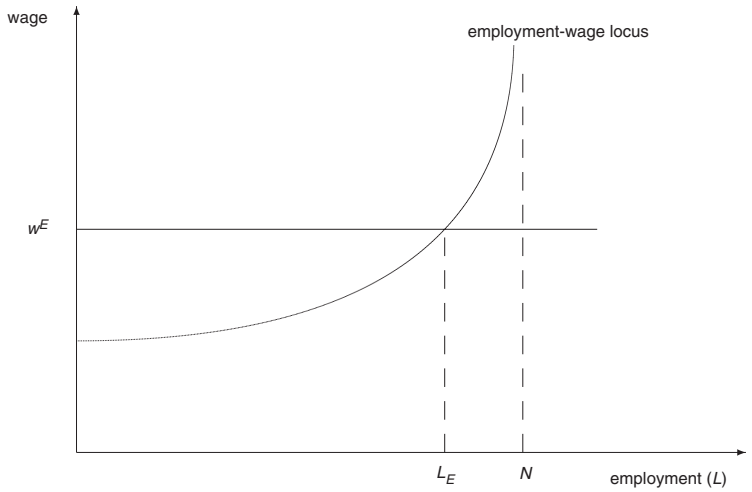


Figure 1 ***Equilibrium wage and employment.***

so in turn we can use (17) and (20) to solve for V^u and substitute into Eq. (19), giving

$$w = e^* + \frac{1}{\phi(z^*)}\left(a + \frac{\phi(z^*)^2}{\phi'(z^*)}\right)\sigma. \tag{21}$$

Equation (21) gives the locus of potential equilibrium values of w and a.

Now job loss among the employed occurs with probability $F^* \equiv F(\bar{x}^*, e^*)$, while job finding among the unemployed occurs with probability a. So if u is the steady state unemployment rate, we must have $(1-u)F^* = ua$. Solving for a and substituting into (21), and substituting also for F^* using Footnote 18, gives

$$w = e^* + \frac{1}{\phi(z^*)}\left(\rho + \frac{F^*}{u}\right)\sigma. \tag{22}$$

This expression shows potential equilibrium wage and unemployment levels for the labor market.

Figure 1 shows the market equilibrium when long-run labor demand is perfectly elastic, at w^E, with N indicating available workers and L_E indicating employment. (More general formulations are easily handled.) Equilibrium unemployment, u^E, solves

$$w^E = e^* + \frac{1}{\phi(z^*)}\left(\rho + \frac{F^*}{u^E}\right)\sigma. \tag{23}$$

Clearly $u^E > 0$. Also, inspection of (23) shows that an increase in σ increases unemployment. This outcome is intuitive. The weaker the link between the dismissal threat and employee behaviors, the stronger are the incentives required to elicit the desired effort level. In equilibrium, heightened incentives require higher unemployment.

This model of equilibrium unemployment—the Shapiro-Stiglitz model—has emerged as a workhorse for the analysis of macro-labor issues. It has proved to be useful also for evaluating policies like unemployment benefits, the public interest in regulating firms' layoff decisions (i.e., just-cause dismissal requirements, as discussed in Levine, 1991), and the potential of minimum wage policy to actually increase employment (Rebitzer and Taylor, 1995b). Having said this, economists are divided on the extent to which efficiency wages are an important source of equilibrium unemployment. Other forces, like labor market frictions, matter as well in determining equilibrium unemployment rates.[20] Efficiency wages are clearly not the whole story.

Labor market segmentation

Although we have focused on unemployment, the lost output from unemployment may not capture the full social costs of efficiency wage personnel policies. After all, if there are some jobs in the labor market where agency issues are of little importance, workers should generally be able to find jobs there. From an efficiency perspective, finding work in these "secondary jobs" is similar to unemployment in that individuals in the secondary labor market would prefer higher-productivity "primary jobs," but the equilibrium supply of qualified workers for these jobs exceeds the demand.

Labor market segmentation emerges if we enrich our efficiency wage model by allowing the difficulty of agency problems to vary across firms. Recall from (21), that in the Ritter-Taylor version of the Shapiro-Stiglitz model firms choose to pay

$$w = A + B\sigma, \tag{24}$$

where A and B are positive constants that are independent of σ. Firms that have low values of σ, i.e., who have production processes with accurate measures of worker effort, can pay wages that are relatively low. On the other hand, firms will choose to set wages high when effort is hard monitor or, equivalently, when they face high values of σ.

This latter observation was emphasized in Bulow and Summers' (1986) paper on "dual labor markets." In their conception, firms with severe agency problems pay high wages and are said to belong in the primary sector. The strategy of paying high wages is effective because workers are motivated by the prospect of retaining valuable jobs. Thus we also expect to observe low levels of voluntary exit from such firms and

[20] Hornstein et al. (2007) offer a review and discussion of models of unemployment resulting from search frictions. Search models produce both unemployment and wage dispersion, but search frictions sufficient to account for equilibrium unemployment imply far less wage inequality than is actually observed.

efforts on the part of firms to retain workers even in a down-turn. In contrast to the primary sector, firms that have modest agency problems can pay wages that are close to the market-clearing level. These secondary sector firms will be less concerned about worker turnover. In an extension of this argument (Rebitzer and Taylor, 1991) we show that firms which employ efficiency wages as a motivating device will also be led to hoard labor, i.e., employ labor above the value-of-marginal-product curve. By taking actions to ensure future employment—perhaps by hiring contingent workers to absorb demand shocks—firms can reduce the wage needed to provide optimal motivation to workers.

The most widely examined empirical prediction of efficiency wage models of labor market segmentation is that there will be cross-firm and cross-industry wage variation resulting from firm characteristics, rather than worker characteristics. There is considerable evidence for industry and firm wage effects (including well-known work by Krueger and Summers (1988), on industry effects, and Brown and Medoff (1989), on firm size effects) but it is often unclear how much this is due to factors emphasized in efficiency wage models (such as monitoring difficulties) or other market imperfections such as those emerging from search frictions (e.g., Burdett and Mortensen, 1998).

One potentially helpful approach entails the study of specific firms and industries, with an eye toward the central predictions of the model. Thus, Cappelli and Chauvin (1991) examined worker performance across plants within the same firm, examining the extent to which workers seem to choose performance on the basis of the value of their job relative to other opportunities in their local labor market. They find evidence that is generally supportive of the efficiency wage set-up. Similarly, in work with Daniel Nagin and Seth Sanders, (Nagin et al., 2002), we evaluated a field experiment in which a firm manipulated monitoring rates across several work sites. Consistent with the effort-regulation model set up above, there was substantially more malfeasance in locations with low monitoring levels.

At the broadest level, the efficiency wage literature points to important social costs that emerge as a result of the strategies individual firms use to resolve agency problems. If firms indeed rely on the fear of job loss to motivate employees, labor markets can be expected to waste human capital on a large scale through involuntary unemployment and labor market segmentation. If, however, other motivators can be mobilized to resolve agency problems, the situation may not be so grim. The costs of agency problems might be further reduced if schools can socialize children to be especially responsive to these alternative motivators. Indeed, some have speculated that such socialization may be the source of much of the social and private returns to investments in human capital. We take up some of these alternative motivators in Section 4 below. Before turning to these, however, we must first introduce another conceptual building block that is important for our story—incentives that are intended to work along more than one dimension.

3. EXTRINSIC REWARDS AND DUAL-PURPOSE INCENTIVES

In real-world applications, compensation policies are often asked to do "double duty." A well known and intuitive example of this is Lazear's (2000) study of compensation practices at Safelite, a windshield installation company. As might be expected from the basic principal agent model, the introduction of an explicit piece rate system induced many workers to perform at a higher intensity level. In addition, the piece rate system had a *selection* effect: workers who disliked having to choose between lower compensation and a faster pace of work left the firm while, at the same time, the firm was able to attract workers drawn to the income-effort tradeoffs inherent in the piece rate system. In this case, incentive pay was serving a dual role: motivating and attracting employees.

At Safelite, the selection reinforced the effect of incentives on work effort. In many cases, however, there is a tension between the multiple effects of incentive pay, and thus employers must make compromises along one dimension in order to accomplish an objective along a second dimension. In the following three sub-sections we give examples of this phenomena and illustrate how the introduction of double duty incentives helps address well known anomalies. The discussion in this section also sets up of the discussion of behavioral models that follow. The special problems posed by dual purpose extrinsic incentives can be either ameliorated or sharpened by behavioral factors of the sort discussed in Section 4. In addition, dual purpose incentives play a key role in the models of intrinsic motivation presented in Section 5.

3.1. High wages as a signal of firm fitness

We begin by discussing a theoretical issue that is well known in the literature on efficiency wages. As we have seen, firms that pay efficiency wages must set wages above the market clearing level to elicit the desired level of effort. Effort can, however, be elicited more cheaply by a deferred compensation policy that causes employees to, in effect, *post a performance bond*. By judiciously back-loading pay, firms can create powerful work incentives while choosing a wage path whose discounted present value is equal to the market clearing wage. With this option available, why would employers ever select a more costly efficiency wage strategy? The practical relevance of this theoretical puzzle is sharpened by empirical work suggesting that even when very large amounts of deferred compensation were available, as is the case in the promotion from associate to partner in large corporate law firms, firms set wages as if they were pursuing an efficiency wage strategy (Rebitzer and Taylor, 1995a).

Ritter and Taylor (1994) tackled this issue by observing that for both efficiency wages and the performance-bond incentive, the power to shape behavior depends on the likelihood that the firm will honor its future commitments to employees. All else equal, firms will more effectively solve agency problems if employees expect them to be highly reliable in honoring future wage commitments.

Ritter and Taylor build upon this insight by positing a market in which there are two types of firms: highly reliable firms (i.e., firms that are unlikely to go bankrupt or otherwise renege on commitments) and less reliable firms (firms that are more likely to become bankrupt or renege). Reliability is known by the firms but not by anyone else, though the distribution of types is common knowledge. Firms would like to resolve their agency problem as cheaply as possible, and are inclined to do so by asking workers to post bonds in the form of deferred compensation. The posted bond is forfeited if the worker is judged to be working at a sub-standard effort level but is returned, with interest, if the worker's observed performance meets the expected standard.

Under some conditions, all firms pursue the same deferred compensation strategy. In this *pooling equilibrium,* workers will require a rate of return on their bonds that reflects the aggregate level of riskiness, based on the market-wide probability a firm will fail and be unable to return the bond.[21] This, of course, is a good deal for low-reliability firms—who benefit by paying a below-market interest rate on the bonds that workers have posted—but a bad deal for high-reliability firms. A more interesting possibility is that efficiency wage strategies might emerge for some firms as a *separating equilibrium*. In this equilibrium, a reliable firm that deviates from the bonding strategy—by paying a high wage up front—offers a credible *signal* that it is a highly desirable counterparty for long-term relationships. If the offered wage is sufficiently high, low-reliability firms will find it unprofitable to mimic this strategy, and the equilibrium thus satisfies the "Intuitive Criterion" of Cho and Kreps (1987). Equilibrium efficiency wages arise endogenously, in short, without a recourse to limited liability arguments.[22]

Our primary point here concerns the use of incentives policies to do "double duty." In the separating equilibrium, highly reliable firms use wage policies to solve an agency problem *and* to signal the fitness of the firm. In order to pursue both objectives, these firms must compromise on their use of deferred compensation and this compromise necessarily introduces distortion. Thus, in the separating equilibrium there is a surplus transfer to workers employed by highly reliable firms and employment in the reliable-firm sector is inefficiently low.

3.2. The rat race

Rat race models build on a simple observation: early in their careers many successful professionals appear to be overworking. It is commonplace to find lawyers, consultants, and assistant professors complaining that the hours they work are simply "too much" and that they interfere with forming and raising a family. These strains are increased by

21 In this model, young workers have concerns about the realization of high earnings at the firm later in their careers. Wages paid to young workers thus depend on the degree to which firms are judged to be unstable.

22 In the law firm context, the term of art for paying very high wages to summer interns and associates is "paying full freight." Law firms that are able and willing to "pay full freight" signal that the value of their partnership is high, and this in turn allows them to attract the best talent.

the dramatic influx of women into professional occupations because overwork is most intense during prime years for family formation and childbearing. From the point of view of simple models of labor markets, this sort of overwork is anomalous. Firms are in competition for talent, and it would seem that the most successful competitors would be those who best accommodate employee preferences about work conditions—including work hours. In his famous paper on the "rat race," Akerlof (1976) offers a potential resolution to this anomaly based on unobservable worker heterogeneity.

Akerlof's set-up focuses on a production line. At the end of the day, line workers are jointly rewarded on the basis of total output. There are two types of workers—those inclined to work hard and those inclined to work less hard. To employers these workers appear identical, so they all earn the same wage. If both types of workers accept positions on the production line, this is a great deal for low-effort workers (who would earn lower pay than high-effort workers in a perfect-information world) but a bad deal for high-effort workers. This is precisely the situation that might lead firms to adopt rules that will provide high-effort workers the opportunity to credibly *signal* that they are in fact high-effort workers. Thus, Akerlof's proposed solution is that the firm set the production line at a speed that is uncomfortably fast for high-effort workers but more uncomfortable yet for low-effort workers—so uncomfortable indeed that the low-effort workers will opt out of working for the firm. The rat race thereby serves the useful function of screening out the low-effort workers.

In Akerlof's model, compensation policies are, quite clearly, doing "double duty." Compensation arrangements and work conditions are structured (i) to compensate workers at a level necessary to induce them to accept employment at the firm, and (ii) to create incentives that attract the "right kind" of worker to the firm. The distortion here is that workers are being asking to provide effort that exceeds the first-best level. In a marketplace with many employers, the market can devolve into an equilibrium in which *all* firms that hire high-effort workers ask those workers to work at uncomfortable effort levels. This "adverse selection" equilibrium occurs because any one firm failing to adopt a rat race would be swamped by low-effort workers. The equilibrium is inefficient, in the sense that all firms would experience increased profitability if they coordinated on a lower-effort work norm.[23]

Akerlof's demonstration of an overwork equilibrium was presented in a self-consciously unrealistic example, but subsequent theoretical and empirical work suggests that it is an important phenomenon in professional labor markets. For instance, in a paper with Renee Landers (Landers et al., 1996), we embed Akerlof's idea into a simple tournament-partnership model designed to shed light on work practices in

[23] Intuitively, the over-work equilibrium might persist even when there is a small number of low-effort workers. No one firm can deviate from the equilibrium without suffering harm from adverse selection. But if all firms backed away from overwork requirements, any one firm would get stuck with only a negligible number of low-effort workers. All firms would be better off.

large US law firms. In our two-period model, young lawyers accept salaried positions as "associates" for one period, and if deemed suitable are promoted to be "partners" (equity shareholders) in the subsequent period. Partners share equally in firm surplus. This equal sharing rule gives incumbent partners powerful incentives to promote only highly motivated lawyers into the partnership.

We assume that there are two types of lawyers who are equally productive but have differing preferences over the hours they prefer to work: these are "short-hour" and "long-hour" workers. Now in our setting, firms have the incentive to attract workers who will be inclined to work long hours. The reason is that when workers become partners—at which point they share firm surplus—the long-hours individuals will engage in less free riding. As in Akerlof's model, an adverse selection equilibrium emerges. In our case, associates' willingness to put in extended hours over many years serves as a credible signal that they are long-hour individuals.

Empirical evidence for the relevance of the over-work equilibrium comes from an empirical examination of work hours and work preferences in associates at two large East Coast law firms. In a survey conducted in these firms, we find that most associates express a preference for working shorter hours (with a correspondingly lower salary) but, importantly, their willingness to work shorter hours hinges on the work-hour norms adopted by other associates. In addition, we find that in making promotion decisions, partners use an associate's willingness to work long hours as an indicator of the motivation associates have to excel. These findings would not be expected in a conventional labor market, but are precisely what one would expect if work hours are being used as a signal of one's otherwise-unobservable type.

In our paper (Landers et al., 1996) we abstract from "career concerns" outside one's own firm, but it is clear that overwork early in one's career might be valuable not only as a signal within a firm but as a means of career advancement elsewhere. Completion of six years as an associate at a law firm known for abusing associates with grueling hours can be a valuable means of demonstrating an important but hard-to-observe trait to other employers in the marketplace.

The key idea that current on-the-job behaviors can affect one's future career, through their impact on reputation, is studied in insightful papers by Holmström (1999) and Gibbons and Waldman (1999). Gicheva (2009) shows how long work hours early in one's career can affect an individual's value in the market later in the career. Gicheva shows, further, that her model helps explain wage growth in a sample of workers who took the Graduate Management Admissions Test (GMAT). Specifically, she shows that among workers who worked above-norm hours when they were young (48 or more hours per week), subsequent wage growth was positively correlated with early career hours worked. The same was *not* true for those workers who worked fewer than 48 hours; for those workers wage growth was uncorrelated with hours worked.

Signalling can be particularly dysfunctional in situations in which workers can devote effort to more than one task—an issue we take up next—because it can distort effort allocation. An example is given in the work of Acemoglu et al. (2008). In their model, career concerns can motivate excessive and misguided signaling by primary school teachers—misguided because some effort is devoted to improving the signal (student performance on a proficiency test), without actually improving students' true human capital. An important conclusion in that paper is that the problem of excessive signalling can fundamentally shape the desirability of using markets *versus* the government for the provision of some services. Specifically, in a competitive market, there will be socially costly distortions as the result of excessive signalling. An advantage of having teachers employed in the public sector is that the government might be able to commit to policies that eliminate excessive signalling.

3.3. Multi-tasking

Perhaps the most obvious case of dual-purpose incentives occurs when a principal seeks to regulate an agent's behaviors along more than one dimension. We have already encountered examples along these lines in our discussions above. For instance, in our examination of CEO compensation, we noted problems that arise when a compensation board seeks to create incentives for a CEO to exert effort toward increasing shareholder value *and* limiting wasteful expenditure in the executive suite. Our discussion of tournament incentives suggests a second obvious example: What happens in a tournament when each worker must be motivated to provide effort along his own assigned task *and* be motivated also to be cooperative with other workers?[24] A third example is Acemoglu, Kremer, and Mian's work, mentioned in the previous paragraph, in which teachers allocate effort that improves student human capital *and* effort that merely improves a student's test score ("teaching to the test").

Holmström and Milgrom (1991) establish a number of insightful and surprising results in precisely such contexts. The central point of their paper is both simple and profound: when an agent performs multiple tasks, incentives must perform the double duty of inducing appropriately high levels of effort generally *and* inducing a desirable allocation of an agent's attention across the various tasks inherent in the job.

We can get a feel for their analysis by making a simple extension to our baseline principal agent model. Let us suppose now that the agent can allocate effort along two dimensions, $e_1 \geq 0$ and $e_2 \geq 0$. We suppose also that the agent's utility is now $w + d(e_1 + e_2)$, and, following Holmström and Milgrom, we make two key assumption about the function that gives money metric disutility of effort, $d(\,)$: (i) it is convex and (ii) it achieves a maximum at a positive level of effort. This last assumption means that in the absence of direct incentives the worker will be happiest when putting forth some

[24] A number of papers have taken up this issue, including Lazear (1989).

effort. The principal's objective continues to be the maximization of value added by the worker, now given by $g(e_1, e_2) - w$.

A principal who has reasonably accurate measures of e_1 and e_2, say x_1 and x_2, will be able to construct an incentive scheme in which bonuses reward each dimension of effort appropriately. Matters are more interesting when the principal has good information along one dimension of effort but not the other. To take an extreme example, suppose the firm has a subjective measure x_1, but no measure at all of x_2. The firm's best strategy then will depend crucially on the nature of the production function, $g(e_1, e_2)$. To see how this works, we set $d(e_1 + e_2)$ to be $-\frac{1}{2}(e_1 + e_2 - e_B)^2$, with $e_B > 0$ representing the agent's "bliss" level of effort, and work out the optimal incentives for two different production functions: first, a case in which the two types of effort are perfect substitutes, $g(e_1, e_2) = a_1e_1 + a_2e_2$; second, a case in which they are complements, $g(e_1, e_2) = e_1e_2$.

The solution to the first case is easy to characterize. If the firm places any incentive whatsoever on the first type of effort (i.e., a bonus based on x_1), the worker's best response will be $\hat{e}_2 = 0$, and with this in mind the firm can follow the steps outlined for our baseline one-dimensional principal agent model. It is easy to confirm that the result will be to elicit effort $e_1^* = a_1 + e_B$. Intuitively, the firm will prefer this strategy if the value of the first type of effort is high relative to the second type of effort, which is certainly true if $a_1 \geq a_2$. On the other hand, if a_2 is sufficiently large, the principal might decide to use no explicit incentive and instead ask (nicely!) that the worker direct all his effort along the second dimension. Given a binding participation constraint, $w - \frac{1}{2}(e_1 + e_2 - e_B)^2 = v$, it is easily shown that value added by the worker is

$$g(e_1, e_2) - w = \begin{cases} a_1e_B + \frac{1}{2}a_1^2 - v & \text{when incentives are placed on } e_1, \text{ and} \\ a_2e_B - v & \text{when agent effort is directed to } e_2. \end{cases} \tag{25}$$

As anticipated, if a_1 is sufficiently high relative to a_2, the firm will simply place incentives on the observable portion of performance. This is efficient when $a_1 \geq a_2$, and is second-best optimal even when a_1 is moderately lower than a_2. However, when a_1 is sufficiently low relative to a_2, the firm will instead try a "cooperative" strategy. No incentives are used; the worker is simply asked to direct all effort to the second dimension.

The solution to the second case, in which the two types of effort are complements, is also intuitive. When $g(e_1, e_2) = e_1e_2$, the firm clearly must avoid a best response of $\hat{e}_2 = 0$, and so will *never* use an explicit incentive along the first effort dimension. In this case the firm instead directs the worker effort to be $e_1 = e_2 = \frac{e_B}{2}$ and hopes that the worker complies.

Simple as this analysis is, several interesting points emerge:

First, we see that there will be cases in which an employer will choose not to place an incentive on an easily-observed dimension of performance, even when that effort is

valuable to the firm. This happens when there is a similarly-valuable second dimension of effort that is sufficiently difficult to observe and incentivize. Such an outcome is particularly likely when multiple tasks are complementary. In such cases the firm is best off using very low-powered incentives, i.e., simply paying a base wage.

Second, as shown by our second example, the principal's optimal incentive plan can result in a second-best outcome that is far from efficient. For instance, when $g(e_1, e_2) = e_1 e_2$, it is easily confirmed that the efficient level of output is e_B^2.[25] The firm's low-powered incentive scheme results, instead, in output $\frac{1}{4}e_B^2$. If possible the firm would very much like to find a way to make this worker a residual claimant, and indeed would be willing to suffer substantial cost along some other dimension to make this happen. In short, a strong motive exists here to outsource the task at hand to an independent contractor if this can be made workable.[26]

Third, when a firm cannot use independent contracting, or for some compelling reason chooses not to, a central goal of the firm often entails structuring activities to improve management's ability to closely monitor and supervise key contributions by employees. Put another way, the issue of multi-tasking can matter a great deal for the organization of firm production.

Finally, and most important for our purposes, the multi-tasking approach developed by Holmström and Milgrom, and illustrated with the simple example above, clearly opens the door for "behavioral factors" to play a central role. For example, in our model, the "bliss" level of effort (e_B) is taken to be an exogenous constant. This unfortunate abstraction overlooks the many sociological and psychological factors that determine how intrinsically motivated individuals contribute to the success of their firm. These are the sorts of consideration that lead one into the territory of behavioral economics.

4. BEHAVIORAL APPROACHES TO AGENCY AND MOTIVATION

In this section of the paper, we expand the psychological and sociological foundations of agency models. Our focus will be on behaviors that seem to us especially relevant to the understanding of agency—"social" or "other regarding" preferences.

[25] Maximize $e_1 e_2 - w$ subject to $w - \frac{1}{2}(e_1 + e_2 - e_B)^2 \geq v$.

[26] From a legal perspective, employees are distinguished from independent contractors by the extent of control and supervision the principal exerts over the actions of the agent. A large literature focuses on the forces that drive firm boundaries, focusing on such issues as the direction of employee activities (e.g., Coase, 1937 and Simon, 1951), and firm ownership of assets (e.g., Williamson, 1985; Grossman and Hart, 1986). The relationship between these issues, and the agency problem—particularly in the multi-tasking setting—is developed clearly in Holmström and Milgrom (1994).

Many papers examine the relationship between firm boundaries and employment relationships in specific industry settings. See, as examples, Arlen and MacLeod's (2005) analysis of physicians in managed care organizations, and the research one of us conducted on worker contracts in the petrochemical industry (Rebitzer, 1995).

Social preferences arise because people are naturally inclined to compare their own payoffs, sacrifices and behaviors to others, and often care about the impact of their actions on others. Economists have long understood that these "other regarding" preferences are important.[27] Recent progress in the behavioral economics literature has greatly deepened this understanding through the development of new theoretical models and novel empirical investigations using both experimental and observational data. As we discuss below, social preferences matter for agency problems because they offer an explanation for the norms and reference points that agents use to assess their pay, work effort and happiness.[28] An important insight from the behavioral approach is that the role played by these norms and reference points varies depending on whether one is considering the agency problem in isolation or in a competitive setting.

Our discussion of social preferences in agency considers four distinct but related manifestations of other regarding preferences: pay status, effort norms, professional norms and identity. In each of these sections we begin by sketching a model that makes modest modifications to the standard agency models discussed above. We then consider how well the central predictions of the enhanced model are supported by available empirical analysis.[29]

4.1. Pay status: Financial incentives and inequality aversion within firms

People dislike inequality—especially when they have drawn the short straw. Indeed, there is substantial indirect evidence that wellbeing is shaped in large measure by comparisons

[27] For example, several decades back Gary Becker initiated important strands of inquiry in economics by positing preferences that incorporate such factors as "altruism" (Becker, 1981) or "distaste" for interaction with people of a different race (Becker, 1957).

[28] Assessments based on reference points play an important role in behavioral economics generally, included features germane to labor economics. Kahneman and Tversky's "prospect theory" of decision making under uncertainty argues that individuals are loss averse and that they calculate gains and losses relative to (potentially manipulable) reference points (Rabin and Thaler, 2001). Sometimes reference points are determined by the status quo or by inertia (e.g., Thaler and Sunstein, 2008; Genesove and Mayer, 2001).

We do not discuss labor supply here, but note that reference points can be important in those models as well. Camerer et al. (1997), for instance, argue that the labor supply of taxi drivers seems to entail drivers evaluating their daily income relative to a daily target. (See also Farber (2005, 2008) for additional evidence, some of it to the contrary, and DellaVigna (2009), for a clarifying discussion.) Fehr and Goette (2007) provide evidence for a field experiment suggesting that loss aversion and reference points may be important in determining work intensity.

[29] Camerer and Loewenstein (2004) provide a nice justification for this approach: "Theories in behavioral economics . . . strive for generality—e.g., by adding only one or two parameters to standard models. Particular parameter values then often reduce the behavioral model to the standard one, and the behavioral model can be pitted against the standard model by estimating parameter values. Once parameter values are pinned down, the behavioral model can be applied just as widely as the standard one."

More generally, Camerer and Lowenstein's paper provides an insightful introduction to a broad and rich set of ideas in behavioral economics, including observations about the origins of modern behavioral economics, and suggestions about future directions for the field. DellaVigna (2009) gives a good recent review of behavioral economics, focusing on evidence drawn from the field.

with others.[30] This idea can matter within organizations because people are likely to compare themselves with others around them in the workplace. In turn this can be an important determinant in shaping firm compensation policies.

The idea that interpersonal comparisons matter to agents can easily be captured by including "asymmetric inequality aversion" into utility functions. Utility is increasing in one's own income, of course, but decreasing in the income of other relatively-wealthier comparison individuals. The "asymmetry" refers to a presumption that agents suffer more from inequality that is to their material disadvantage than they gain from inequality that is to their material advantage (see, e.g., Fehr and Schmidt, 1999).[31]

As an example of how asymmetric inequality aversion can affect our agency models, recall the tournament model, as set out in Section 2.2.2. In that model, workers had utility given by $w-e$, and made a net contribution of $g(e)-w$. The firm conditioned pay on an imperfect measure of e: it paid a base wage w_0 to all workers, and in addition gave a bonus b to the fraction $(1-P)$ of workers who had the highest observed performance. (The "bonus" in this case would typically be a promotion to a higher-paying job.) Given workers' best-response effort choices, we saw that this simple tournament resulting in all workers supplying the efficient effort level, i.e., they choose $e^* = \hat{e}(b^*)$ that solves $g'(e^*) = 1$.

Suppose we take that same model but now introduce asymmetric inequality aversion by letting

$$\text{utility} = \begin{cases} w_0 + b - \delta_W b - e & \text{for "winners," and} \\ w_0 - \delta_L b - e & \text{for "losers."} \end{cases} \tag{26}$$

Here, $\delta_W \geq 0$ reflects the possibility that winners feel empathy for losers, proportional to the inequality generated (but of course people do like to win, so $\delta_W < 1$). $\delta_L > \delta_W$ reflects the fact that workers who do *not* win the bonus suffer an even large utility loss due to inequality aversion. Repeating steps outlined in Section 2.2.2, we can show that the principal's solution now has a first-order condition

$$[g'(\hat{e}(b^{**})) - 1]\hat{e}'(b^{**}) = (\delta_L - \delta_W)P > 0, \tag{27}$$

[30] This is the basis, for example, of the well-known "Easterlin paradox"—the paradoxical results that (i) individuals who are low in a nation's income distribution report themselves to be unhappy, (ii) the average level of unhappiness does not much vary across nations with different levels of aggregate income, and (iii) countries do not get much happier as they get richer. While there is evidence to suggest that absolute income matters for happiness (e.g., Stevenson and Wolfers, 2008), it seems clear that one's standing on the economic totem pole matters as well. Frank's (1985) well known book provides an interesting and wide-ranging discussion on the human inclination for social comparison.

[31] The assertion that utility is influenced by inequality aversion represents a "stripped down" way of characterizing the behavioral phenomena under study. Our initial treatment makes no distinction about agent attributions concerning the nature of inequality (e.g., what the inequality might say about the principal's intentions or other agents' intentions). We consider more sophisticated approaches below.

and, since $e'(b) > 0$, we have $g'(e^{**}) > 1$, which in turn means that the firm here settles for a second-best effort level, $e^{**} < e^*$.

Inequality aversion causes the incentive pay parameter to do the "double duty" of eliciting work effort and determining the extent of expected pay inequality in the firm. As a result, the firm must compromise along an important dimension by lowering incentive pay, b, and reducing the effort level elicited from workers. Equation (27) also shows that in this setting the firm will want to be careful how it sets its promotion rate. Here the firm would like (all else equal) to set P near 0, which would allow effort to approach first-best. Intuitively, the cost to the firm of inequality is lowest when there are relatively few people who are affected by the inequality, i.e., when the promotion rate, $1 - P$, is close to 1.

The logic of this model of income comparisons underlies Frank's (1984) seminal article on inequality aversion in labor markets. In his treatment, an employee gains in utility from being high in the firm's pay hierarchy and loses utility from having a low position. Just as in our model, these concerns cause firms to operate with lower-powered incentives. Frank presents evidence drawn from many different types of organizations; he sees, for example, a dampening effect in commissions paid to car salesman and realtors as well as pay compression among college professors.

Encinosa et al. (2007) present a similar analysis of these issues in the context of incentive pay within medical partnerships. In these professional organizations, physicians determine incentive intensity by choosing how broadly they wish to share the income they generate with others in the practice. For example, groups often choose to share income equally across physicians—in a practice with n physicians, each physician keeps $1/n$ of profits—which minimizes inequality. The practice of equal sharing rules has the potential disadvantage, though, of offering the lowest possible level of incentive intensity to partners. The model of inequality aversion set out by Encinosa, et al. shows that the tension between these forces makes sharing rules less attractive in large partnerships than in smaller partnerships—a result supported by the available data. The authors also present evidence consistent the notion that physicians compare effort as well as income. We take up effort comparisons in the next section.

As a second example of the potential impact of inequality aversion in a principal agent setting, consider the multi-tasking model we examined in Section 3.3. Recall that in that model, we assumed an agent's utility is represented by $w - \frac{1}{2}(e_1 + e_2 - e_B)^2$, where e_B is a positive constant, and we assumed further that the principal had a good signal for e_1 but not e_2. So when the principal's payoff is $e_1 e_2$, the best the principal can do is pay the agent a fixed wage w that meets the participation constraint (i.e., so that $w = v$) and then direct the agent to allocate effort so that $e_1 = e_2 = \frac{e_B}{2}$. Now suppose that the agent is inequality averse, just as in (26), but that in this case her comparison is the *principal's* income, π. So for the agent, utility is

$$w - \frac{1}{2}(e_1 + e_2 - e_B)^2 - \delta(\pi - w), \quad \text{if } w < \pi, \tag{28}$$

where $\delta > 0$ reflects the extent to which the agent is inequality averse. The important point here is that the agent can always costlessly enforce perfect equality by simply adjusting effort allocation (keeping total effort at e_B); she can reduce the principal's profits, while causing no harm to herself. If the principal sets the wage to v, the agent will adjust effort so that π also equals v. So the principal typically finds it profitable to increase w above the participation constraint, i.e., the principal uses "rent sharing." Reducing wages to the level of the employee's outside option would be self-defeating in this context because it risks that the agent will become disgruntled and take steps to "even the score."

The idea that inequity aversion supports rent sharing has been extensively explored in laboratory and field experiments involving variations on the Ultimatum Game. In this bargaining game a *proposer* offers to divide a fixed amount of money between himself and a responder. The *responder* can accept or reject this offer. If the offer is accepted, the money is divided according to the offer. If the offer is rejected by the responder, however, neither the responder nor the proposer gets any money. The conventional game theoretic solution to this bargaining problem for selfish players is for the proposer to make an offer in which he keeps all (or nearly all) of the money while the responder accepts any offer.

It turns out, however, that the participants in these games don't behave as entirely selfish players. Proposers routinely make offers close to an equal division of the pie and responders routinely reject low offers. These anomalies can be resolved, of course, by introducing equity concerns into players' utility functions. Inequality averse proposers get less utility than entirely selfish players do by proposing a division of the pie that greatly favors themselves. Conversely, inequality averse responders can credibly threaten to destroy the surplus if highly unequal divisions are proposed. In practice, reasonably egalitarian offers are made and accepted.

Most employment relationships exist in the context of labor markets. Thus it is not sufficient to demonstrate that individuals prefer more equitable pay practices. Economists must also establish that these preferences matter for the equilibria that emerge in markets. Fehr and Schmidt (1999) examine this central issue by considering whether the egalitarian rent sharing observed in the Ultimatum Game survives in an environment in which there are many proposers and a single responder who must accept or reject the best offer received. They find that competition between proposers leads to lower levels of egalitarianism. To see the logic, consider a situation in which n proposers each offered 50 percent to the responder, leaving each proposer with a $1/n$ chance of having his offer accepted. An individual proposer could clearly do better by simply offering a 51 percent share to the responder, thereby insuring that his offer was selected (the probability increases from $1/n$ to 1). But all proposers are driven by this same logic, and in the end

the responder gets all the surplus. This, of course, is the outcome we would observe if proposers had no inequality aversion.

The irrelevance of inequality aversion stems from the fact that with many competitors, no single player can prevent an inequitable outcome. If no individual action can reduce the inequality of the equilibrium outcome, then inequality aversion cannot be an important determinant of behavior. Fehr and Schmidt conclude that matters are different when individual players have a way to impose a cost on the counterparty to a highly unequal offer (as does the agent in the example we consider with Eq. (28)). Specifically,

> *…competition renders fairness considerations irrelevant if and only if none of the competing players can punish the monopolist by destroying some of the surplus and enforcing a more equitable outcome. This suggests that fairness plays a smaller role in most markets for goods than in labor markets. This follows from the fact that, in addition to the rejection of low wage offers, workers have some discretion over their work effort. By varying their effort, they can exert a direct impact on the relative material payoff of the employer (Fehr and Schmidt, 1999).*

In short, agency problems of the sort depicted in our model above (that allows retaliation motivated by equity concerns), can survive market competition.

The Fehr-Schmidt conjecture has been examined experimentally by Fischbacher et al. (2009).[32] They find that increasing proposer competition in the Ultimatum Game, by adding extra proposers, causes a large increase in mean accepted offers. Similarly, increasing responder competition causes a reduction in mean accepted offers. Both of these results suggest that competition undercuts the influence of equity norms on bargaining outcomes, although in each case the increase in inequality of outcomes is less than one would predict on the basis of competition alone.

The idea that workers exact retribution upon employers when treated unfairly is supported by Bewley's (1999) extensive qualitative interviews. He found that managers and other labor market participants believe that there is a connection between employee morale and performance. Bewley focuses on the morale effects of cutting wages in recessions: Employers are averse to cutting wages because of the fear of a backlash from employees.

Evidence that the fear of backlash is reasonable emerges from a number of recent studies by Mas and co-authors. In a remarkable study, Mas (2006) finds that when New Jersey police officers lose in final offer arbitration, so that the wage they receive is lower than the requested wage, arrest rates and average sentence lengths decline, while crime reports rise. This evidence is consistent with the idea that workers are less inclined to provide effort when the wage falls below a salient reference wage. Krueger and Mas (2004) report evidence that a long and contentious strike and the hiring of replacement

[32] That paper also provides reference to a large relevant experimental literature.

workers in a Bridgestone/Firestone plant contributed to the production of defective tires. Mas (2008) finds Caterpillar plants that underwent contract disputes experienced reduced workmanship and reduced product quality. In this latter study he estimates that the contract dispute was associated with at least $400 million in lost service flows due to inferior quality.

The papers cited in the previous paragraph study workers in unionized environments. The presence of a union likely facilitates the sort of collective retaliation that punishes employers who take morale lowering actions. The central idea, that perceptions of unfairness can damage effort, is likely to be important outside the union sector. Evidence along these lines appears, for example, in our field experiment with Nagin and Sanders (Nagin et al., 2002), which manipulated monitoring levels at call centers collecting donations for charitable causes. In that study, conducted in a non-union environment, we had access to a direct measure of malfeasance on the part of individual employees as well as direct measures of individual employee perceptions about the employer. The measure of employee malfeasance is the rate at which employees artificially inflated their level of sales in order to earn extra commissions at the expense of the firm. The employer could catch some, but not all, of this activity through costly monitoring of a random sample of calls. When employees worked in centers with very low apparent rates of monitoring, opportunistic behavior increased. However, this increase was observed only for a subset of employees, and, importantly, increased opportunism was most prevalent among workers who had expressed feelings that the employer treated them unfairly, did not care about them, and provided a bad place to work.[33] Employees who perceived themselves to be unfairly treated struck back at the employer (and added to their own income) when the opportunity to do so arose.

Inequality aversion on the part of employees has a number of interesting ancillary predictions about the way employment relationships are organized. Firms that employ people in both low-wage and high-wage occupations must go to some length to be sure that employees in the low-wage occupation do not include the high-wage occupation in their reference group. Failure to make this separation can lead to pressures to either pay employees in the low-wage occupation too much or employees in the high-wage occupation too little. Indeed, it would not be at all surprising to see firms choosing outsourcing to other firms to avoid just these sorts of invidious comparisons.

Similarly, if employees respond to perceived inequities by retaliating along important, but hard-to-monitor dimensions of work effort and quality, firms that engage in highly unequal pay practices ought to seek out ways to reduce the perceived level of inequality. Secrecy regarding pay is a common human resource practice and it obviously makes invidious pay comparisons more difficult. Some companies, such as Walmart and Lincoln

[33] These attitude questions were collected in an anonymous survey of employees conducted before the field study began.

Electric, famously go to great lengths to discourage ostentatious executive perks and the depressing effects on morale they might engender.

For publicly traded corporations the compensation of top executives is a matter of public record. In practice, however, these companies adopt complex and opaque compensation practices that make it difficult to understand exactly how much and in what ways top executives are paid. Bebchuk and Fried (2003) argue that anomalous features of executive compensation—such as the reliance on "at the money" stock options rather than stock grants—are best understood as efforts to camouflage pay and so avoid "outrage" from shareholders, employees, regulators and other interested parties.

Levy and Temin (2007) examine these outrage costs from an institutional and historical perspective. They argue that the Federal government enforced a set of informal yet egalitarian social norms on executive pay from the post World War II era through the early 1970's. These norms were part of a larger set of institutional arrangements that included powerful unions, high minimum wages, and high marginal tax rates for high earners. The actions taken by Reagan administration in the 1980s (notably the firing of the air-traffic controllers) signaled that the Federal government was leaving such decisions as CEO pay strictly to market forces. The degree to which income norms can be shaped by national institutions and economic policy is an important question that would benefit from additional empirical and historical research.

4.2. Effort norms

Fehr and Schmidt's key idea is that equity concerns constrain firm behavior even in competitive labor markets, because of behavioral features in agency problems inherent in employment relationships. If, as appears to be the case, employees respond to unequal or unfair treatment by taking actions that punish the employer, it is a short further step to presume that morale enhancing activities ought to motivate employees to take actions in the interest of their employer. This is the premise of Akerlof's (1982) gift exchange model of efficiency wages.

Akerlof's model plays a central role in behavioral labor economics because it relies on very different sociological and psychological mechanisms than the standard agency approach presented in Section 2 above. Instead of engaging in calculations about the costs and benefits of working harder, employees in Akerlof's model are motivated by norms governing behavior in the exchange of gifts. If the employer pays employees a wage higher than some reference wage, the employee perceives himself to have received a gift from the employer. This gift creates an obligation to give something valuable in return. In the employment context, the obvious way to reciprocate is to provide the firm with more than the minimally acceptable level of work effort and attention.

Akerlof's approach to the problem of agency rests critically on the concept of effort norms, i.e., on the idea that individuals are motivated to provide effort in ways that enable them to conform to their self image or social identity. Decent people, so the reasoning

might go, return kindness with kindness and so, wishing to preserve the self image of decency, the employee responds to a high wage by returning the favor in the form of high effort to the employer.[34]

Experimental investigations suggest that reciprocity of the sort identified in Akerlof can survive in competitive environments. For example, Fehr et al. (1998) report results from a laboratory experiment in which sellers have the opportunity to select quality levels above the minimum level enforceable by buyers. In treatments where sellers have the opportunity to do so, they reciprocate high prices with high quality levels. Anticipating this behavior, buyers profit by offering high prices far in excess of the seller's reservation prices. In treatments where sellers do not have the opportunity to reciprocate, buyers offer lower prices.

If employee effort responds to the perceived "fairness" of wage offers, then policy makers must pay special attention to policies that might shift perceptions of the "fairness" of a wage offer. Policy may be especially likely to affect fairness if individuals care about employer intentions as well as outcomes. A wage of X in the absence of a minimum wage might be perceived to be quite fair because employers could have offered a good deal less but chose not to. If, on the other hand, the minimum wage was set to be X, then the employer might have to offer a wage above X to demonstrate good intentions.

In an important paper, Falk et al. (2006) investigate whether minimum wage laws influence the perceived fairness of wage offers. They set up an experimental labor market in which individual employees (students paid to participate in the experiment) have to decide whether or not to accept a job offered by a firm. Contrary to the conventional self-interest model, but consistent with a fairness-concerns model, individuals had reservation wages significantly above zero. There was also considerable heterogeneity in reservation wages, giving individual firms an upwardly sloping supply curve for labor. Introducing a minimum wage in this labor market had the effect of increasing individual reservation wages considerably—a result consistent with the hypothesis that the perceived intentions of the firm matter in determining fairness. Surprisingly, there appeared to be hysteresis in the effect of minimum wage laws on reservation wages: subjects exposed to the laws after participating in labor markets with no minimum wage laws increased their reservation wages, but subjects who first participated in labor markets with minimum wages did not revise reservation wages downwards when the laws were removed. These results, if they hold outside the laboratory, have implications that extend far beyond the issue of minimum wage laws. Labor market regulations that influence employer scope of action must take into account how these regulations are likely to affect employee perceptions of employer intentions. More provocatively, the hysteresis result also raises the possibility

[34] This idea is modeled in an insightful way in an important paper by Rabin (1993). Charness and Rabin (2002) provide a clear statement of the ideas, and give reference to further literature. They also give compelling evidence from laboratory experiments on reciprocity. For a discussion of additional experimental evidence, see the chapter by Charness and Kuhn in this *Handbook*.

that regulators might not be able to "undo" the effects of policy simply by reversing previous decisions.

In strong form, well-functioning norms can have considerable social value. They can serve to reduce the problems created by agency in many contexts, including employment relationships within firms.

Given their considerable economic value, it is important to understand the social processes that generate and sustain socially valuable effort norms. In a path-breaking economic analysis, Frank (1988) emphasizes the role of emotions in resolving the "commitment problem," i.e., the problem of eliciting a commitment to constructive cooperation. He argues that rational calculation is often not sufficient to sustain cooperation because by the time the misbehavior occurs, the benefit of punishing the bad actor has often already passed. Emotions, in contrast, can be the foundation of much more powerful sanctions because the commitment to follow through on the action is rooted in the primitive reward structure of the brain. Thus, "cross me and you'll never work in this town again" is a weak deterrent when uttered by a rational calculator who may decide after the fact that it is not worth the effort to punish the double-crosser. It is a strong deterrent, however, when uttered by someone who gets visceral satisfaction in carrying out his threat regardless of the cost to himself.

From a psychological perspective, emotions emerge from a genetically determined neurological reward system. The triggers of this reward system, however, are shaped by an intense and costly socialization process that trains individuals to have a "conscience," i.e., to feel strong emotions when they lie, cheat or otherwise disappoint others' expectations. Evolutionary considerations led Frank to expect that these efforts at socialization will not be entirely successful. Society will be composed of a mixture of types: opportunists who take advantage of chances to free ride and reciprocators who will devote resources to monitoring the behavior of their counterparties and cooperate so long as they perceive others doing the same.[35]

The idea that populations contain a mixture of opportunists and reciprocators is supported by experiments involving public good contribution games. Fehr and Gaecher (2000) study such games and contrast two treatments. In the "no punishment" treatment, anonymous individuals are randomly allocated to groups of four and are given the opportunity to make contributions towards a public good. Payoffs are such that the dominant strategy is to make no contributions towards the public good. In the "punishment" treatment a second stage is added which gives each individual the opportunity to punish others by subtracting from their payout. Punishing poor contributors is costly, however,

[35] In a population entirely composed of cooperators there will be little reason to devote resources to monitoring the actions of others. This is an environment in which opportunists will thrive. Conversely, in an environment with many opportunists, cooperators will enjoy an advantage so long as they devote resources to monitoring their counterparts. Evolutionarily stable equilibria will therefore involve a mix of opportunists and cooperators with the latter spending resources seeking to weed out or punish the former. See, e.g., Gintis et al. (2003).

and no one interested in maximizing their monetary payoff will choose to punish after the damage is done.[36] For this reason one would expect the dominant strategy to be one of "no contribution" in both punishment and no-punishment treatments. This prediction is confirmed in the "no punishment" game: average contributions converge to almost complete free-riding. In contrast, in games with the option to punish in the second stage, individuals do make substantial contributions to the public good and these contributions do not fall over time. Consistent with our effort norms model, subjects are more heavily punished the more his or her contribution falls below the average contribution of other group members. Individuals also exhibit heterogeneous tendencies to free-ride and punish. Depending on the definition, the authors estimate that between 20 and 53 percent of subjects in their study were free riders. Heterogeneity in the tendency to behave opportunistically has important implications for labor markets and personnel practices that we explore further in Section 5.

Emotions can support pro-social behavior in ways other than sustaining irrationally high levels of retaliation against defectors. Ekman (2001), for example, argues that emotional states can be read from the facial expressions of individuals. It follows from this that lying and other opportunistic activities that can elicit strong emotions are harder to sustain during face to face interactions. Valley et al. (1998) investigate this hypothesis in a bargaining experiment which requires negotiators to elicit private information about the true value of an underlying asset when the incentives in the experiment do not support revealing this information truthfully. The study finds that face to face negotiations are more likely to reach mutually beneficial solutions than negotiations conducted over the phone or in writing.

The feelings of guilt and shame that support truth telling and honesty are similar to the emotions that support effort norms, and these emotions are generally thought to be strengthened by physical proximity and face to face interactions (Sally, 2002). A nice laboratory experiment by Falk and Ichino (2006) provides evidence along these lines. In particular, in that study the authors observed "peer effects" in which subjects who would otherwise have provided low effort were motivated to increase effort when physically paired with high-productivity workers.[37]

In a remarkable study of cashiers at a national supermarket chain, Mas and Moretti (2009) find that substituting a worker with below average productivity for a worker with

[36] The fact that some individuals will punish opportunists even when it is not in their direct material interest suggests that punishment is supported by psychological reward mechanisms rather than rational calculation. Consistent with this view, recent brain imaging studies taken during an economic experiment involving trust and retaliation suggest that punishment of individuals who violate trust activates a brain region, the Caudate, that is involved in actions motivated by anticipated rewards (deQuervain et al., 2004). High Caudate activation likely reflects the anticipated satisfaction from punishing defectors.

[37] On net, Falk and Ichino estimate a positive impact on output due to these peer effects similar to estimates of peer effects on absenteeism behavior found in Ichino and Maggi's (2000) study of workers in different branches of a large Italian bank.

above average productivity is associated with an increase in the effort of other workers on the same shift. Low productivity workers are especially responsive to the composition of their co-workers and this peer effect occurs only for low productivity workers who are in the line of vision of the high productivity workers. The effect of high output peers on the productivity of others declines with distance and with the frequency of interaction as measured by the degree to which shifts overlap.

If effort norms indeed require close proximity and frequent interactions within a work group, it is natural to ask whether these motives can operate in large organizations. Very little empirical work has focused on this important issue.[38]

Effort norms clearly matter within organizations and work groups, and may have important implications for broader labor markets as well. In the gift exchange model, as first set out by Akerlof (1982), the "gift" that results in the optimal reciprocal responses from agents is a wage that exceeds the market-clearing wage. The consequence is equilibrium unemployment (see also Akerlof and Yellen, 1990). If some firms and industries find it important to use gift exchange as a motivating tool, and others do not, then the gift exchange model can be used to explain "dual labor markets," i.e., to help understand cross-firm and cross-industry wage variation.

As Akerlof and Yellen (1985) and Akerlof et al. (2000) argue, the gift exchange logic—that worker performance depends on a firm's current wage relative to a reference wage and to the unemployment rate—can be a building block for macroeconomic models. In Akerlof et al. (2000), for instance, a reference wage model is combined with an assumption that some principals adopt "near rational" wage setting rules whereby they ignore the effect of inflation on reference wages when inflation rates are sufficiently low. The consequence is a long-run Phillips curve with the property that a modest rate of inflation (approximately 3% in their calibration) is associated with a lower unemployment rate than is either 0 inflation or high inflation.[39]

4.3. Professional norms

In professions such as law and medicine, the principal agent problem takes on a special importance. Professionals are in theory the agents of their clients, but professionals

[38] A nice exception is Knez and Simester (2001). This case study of Continental Airlines finds evidence of the apparent success a firm-wide incentive scheme that paid out a modest sum of money to almost all employees if the airline's aggregate on-time departure statistics cleared a certain threshold. The authors argue that effort norms, enforced by the relatively small and homogeneous ramp and ground crews at each airport, could and did augment the low-powered financial incentives inherent in the bonus plan.

An important question concerns the extent to which norms can persist in cross-functional work groups where the social distance between members of the group might be high. Such work groups—composed of employees with widely different levels of income, status and education—play an important role in the health care system (e.g., in teams that include surgeons and high school educated technicians working together to improve processes). The failure of effort norms and peer pressure to operate in these settings likely contributes to inefficiencies in our health care system (as described, e.g., in Cebul et al. (2008), and other papers referenced therein).

[39] See also Bewley's (2000) comments on the paper, in which he argues that internal wage comparisons within the firm's wage structure can be a key force in shaping macroeconomic outcomes.

enjoy advantages of education, credentials, status and specialized knowledge that make their clients especially vulnerable to exploitation. In order to protect clients from abuse, professions go to great lengths to inculcate norms of professional conduct. This makes professions an especially important venue for analyzing the effect of norms.

In our analysis of physician incentives in a managed care organization with Martin Gaynor (Gaynor et al., 2004), we develop a simple model of professional norms that we adapt here. The model follows the approach used throughout this essay: we modify the agent's utility function, in this instance to include physicians' regard for their patients. We posit, in particular, that the socialization of physicians causes them to experience disutility when they adopt a practice style that delivers medical services that are less than the level that the patient would select for themselves (if they were as well informed as the physician). Think of this level of services, m_B, as the level (measured here in dollars) that results when the physician incorporates a patient's own preferences into his utility function.

We write the utility of a physician treating $i = 1, \ldots, n$ patients as a function of income earned, w, and the deviation of medical services, m_i from the ideal level:

$$w + \sum_{i=1}^{n} \mu_i d(m_i - m_B^i), \tag{29}$$

where $d(\,)$ is a convex function that achieves a maximum when m_i is equal to a subjective "best" levels of care, m_B^i, and the μ_i parameters indicate the weight the physician places on each patient's well-being. Thus physician utility is increasing in both income and services provided when they adopt a practice style with $m_i < m_B^i$. In a fee-for-service environment where insurers don't try to "manage" the care physicians provide, one would expect physicians to deliver care at or close to m_B^i.

Managed care organizations, such as Health Maintenance Organizations (HMOs), often try to influence physician practice styles through the use of financial incentives. The managed care organization we studied, for example, adopted a simple incentive strategy designed to restrict utilization without substantially harming patients: the principal (HMO) offered agents (physicians) a bonus b if total annual medical expenditures fell below a target $\bar{m}$. The probability that a physician's expenditures on behalf of patients fell below this threshold depended of course on the decisions made on behalf of each patient and also random factors. The probability of earning the bonus, given expenditures m_i and target $\bar{m}$, is given by the c.d.f. $F(\bar{m} - \sum m_i)$.

Giving the physician responsibility for the allocation of resources across a panel of patients in this way makes sense when physicians have practice norms of the sort characterized in (29). We can illustrate the idea easily with the case in which disutility is $d(m_i - m_B^i) = -\frac{1}{2}(m_i - m_B^i)^2$. In this case the physician's best response to a policy, b

and $\bar{m}$ is found by maximizing

$$bF\left(\bar{m} - \sum m_i\right) - \frac{\mu_i}{2}(m_i - m_B^i)^2. \tag{30}$$

This leads to the best response function for the treatment of each patient ($i = 1, \ldots, n$):

$$\hat{m}_i(b) = m_i^B - \frac{b}{\mu_i} f\left(\bar{m} - \sum m_i\right). \tag{31}$$

The extrinsic reward induces the physician to conserve resources on behalf of the HMO, and if μ_i is the same across all patients, say μ (i.e., there is no favoritism), the physician does so in a sensible way.[40] Also, it is easy to confirm that if the second order condition holds, an increase in the bonus induces the physician to reduce expenditures on patients,

$$\hat{m}_i'(b) = \frac{f\left(\bar{m} - \sum m_i\right)}{bf'\left(\bar{m} - \sum m_i\right) - \mu_i} < 0. \tag{32}$$

From (31) and (32) we can see that the intrinsic value the physician places on patients (represented by μ_i) governs the level of expenditures chosen for patients, as well as the power of the extrinsic incentive to alter chosen expenditures.

In our empirical analysis of internal records in an HMO (Gaynor et al., 2004), we found results consistent with the prediction in (32); increased incentive intensity led physicians to reduce expenditures on patients. We also found that physicians cut costs most for outpatient and elective procedures, but not at all for inpatient procedures. Consistent with the model, this suggests that physicians cut costs most where the consequences for patient welfare were lowest.

Two obvious implications of this treatment of professional norms and incentives merit mention. First, patients are not necessarily harmed by incentives that impose constraints on physician actions. Indeed, it is clearly in the interests of patients for their physicians to allocate resources in a reasonable way, because ultimately patients pay for misallocation through higher insurance premiums. Second, physician practice norms of the sort specified above serve to protect patients from potential abuses introduced by cost-containment incentives, especially if the internalization of patient utility is allocated evenly across patients.

Given the pivotal role of professional norms in protecting clients, it is important that attention be paid to the ways in which these norms are established and how they might be undermined. A key example of the latter phenomenon is professional "conflict of interest." In health care, drug companies famously used gifts and aggressive marketing to

[40] Indeed, the outcome is potentially efficient.

influence the prescribing activities of physicians (Avorn, 2004). Conflicts of interest arise in other contexts as well. For example, Jackson (2008) observes that in many financial services markets (including the market for health insurance), brokers who represent one side of the transaction are paid by the other side. These arrangements clearly threaten the ability of brokers to represent the interests of their clients.

Economists have devoted relatively little attention to understanding why practices that create such obvious conflicts of interest persist in markets where principals greatly depend on the independent judgement of professionals. An important exception is a provocative set of articles, Dana and Loewenstein (2003), Moore and Lowenstein (2004) and Cain et al. (2005), which argue that even small gifts can trigger a norm of reciprocity that introduces largely unconscious biases into professional judgements.[41] The fact that these biases are unconscious prevents them from inducing the negative feelings that otherwise cause professionals to conform to norms of acceptable behavior. Laboratory experiments suggest that clients who rely on professional judgements do not adequately adjust their interpretation of professional advice even when they are informed that their agents might be biased. Although the evidence for this view of conflict of interest is far from definitive, the implications for the successful resolution of principal agent models in professional settings are both profound and unsettling.[42]

4.4. Identity

Our discussion of professional norms focused on the idea that physicians might experience disutility—perhaps profound feelings of discomfort or anxiety—if they deviate from proscribed behaviors with respect to their clients or patients. This approach to economic sociology is discussed at length in the work of Akerlof and Kranton (2000, 2005) on "identity." Here is the key idea:

> *The term* identity *is used to describe a person's social category—a person is a man or a woman, a black or a white, a manager or a worker. The term identity is also used to describe a person's self-image. It captures how people feel about themselves, as well as how those feeling depend upon their actions. In a model of utility, then, a person's identity describes gains and losses in utility from behavior that conforms or departs from the norms for particular social categories in particular situations.*
>
> *This concept of utility is a break with traditional economics, where utility functions are not situation-dependent, but fixed. In our conception, utility functions can change, because norms of appropriate and inappropriate behavior differ across space and time. Indeed,*

41 For evidence that judgements of professionals (and others) can be shaped by unconscious but self-serving biases, see Babcock et al. (1995) and Babcock and Loewenstein (1997) and references therein.

42 Another largely neglected theme in the economics of professional norms is whether these norms are strengthened or weakened by market competition. Cooper and Rebitzer (2006) argue that competition between HMOs for patients and providers actually magnifies the importance of physician practice norms, and limits the willingness of managed care organizations to control costs via incentive contracts.

> *norms are taught—by parents, teachers, professors, priests, to name just a few. Psychologists say that people can* internalize *norms; the norms become their own and guide their behavior (Akerlof and Kranton, 2005).*

The idea that "category" and "situation" can be fundamental elements in preferences enormously expands the range of principal agent models. For instance, to the extent that identity can be manipulated within an organization, identity-based incentives might substitute for extrinsic rewards.[43] Just as families and religious communities undertake important and costly investments to ensure that their children internalize a set of values and practices consistent with passing on the family or group's social identity, so organizations might make investments in practices that persuade employees to adopt goals of the organization, and so mitigate agency problems. These investments are likely to be greatest where financial rewards are most costly to the organization, e.g., when performance measures are especially noisy and where high effort (or high effort at peak times) is critical to the organization's success. Investment in identity incentives will also be greatest where inculcating identity is cheap, and it is reasonable to suppose that imparting identity is cheapest when agents are young and/or when highly motivated individuals self-select into the organization—an issue to which we return in Section 5.

The great virtue of identity models is that they are highly flexible and therefore able to account for behavior that is anomalous from the perspective of simple agency models. This virtue is a curse, however, when it comes to generating falsifiable hypotheses for testing identity models themselves. One way around this problem might be to focus on a particular relevant social category and seek to understand key norms that can be studied systematically and characterized in a parsimonious way.

A template for this latter approach can be found in a series of careful and nuanced investigations of psychological factors that generate *gender* differences in economic behaviors. For example, work by Babcock and her co-authors, demonstrates a profound gender-based difference in the inclination to initiate negotiation; "women don't ask."[44] A simple and clear demonstration emerges in an experimental study in which subjects are asked to complete a simple task, and are then put in a position in which there is ambiguity with regard to the payment. In a typical experiment, subjects were told in advance that the payment would be between three and ten dollars. Then, at the conclusion of the session, the experimenter says, "Here's three dollars. Is three dollars okay?" Eight times as many men as women asked for more money in this experiment. Even in a variant of the experiment in which the experimenter provides cues to signal the social acceptability

[43] For example, practices at West Point are designed to "inculcate non-economic motives in the cadets so that they have the same goals as the US Army" (Akerlof and Kranton, 2005), and firm or workgroup loyalty can be found more generally in many organizations.

[44] Babcock and Laschever (2003) provide an engaging and wide-ranging discussion. The authors present real-world evidence about women's general disinclination to ask, and they include observations about the implications for gender inequality.

of negotiation (e.g., with a prompt, "the exact payment is negotiable"), far more men than women take up the opportunity (Small et al., 2007).

In laboratory and field experiments, this disinclination by women to "ask" affects outcomes in negotiated settlements, leading women to do less well than men. Importantly, though, when a woman advocates on behalf of *someone else,* she is typically more successful than when she negotiates for herself, and indeed is generally more effective than men in this capacity (Bowles et al., 2005). Part of the reluctance to "ask," it appears, comes from a desire to avoid self promotion.

Along these same lines, Gneezy et al. (2003) find that women respond differently than men to tournament style incentives when these contests involve both men and women. When paid by piece rate or when competing in single sex tournaments, women's performance is similar to those of men. Niederle and Vesterlund (2007) provide experimental evidence that in comparison to men, women generally shy away from incentives schemes that involve tournament competition.

Gender identity, in short, matters in economically important ways. It is tempting to assert that female identity includes a component that guides women to shy away from competition with men and to reject self promotion. However, it is important to understand that this might not be the whole story, or even the most important part of the story, when using identity to explain gender differences in behavior. There is considerable evidence in psychology that a "kinder, gentler image" is expected of women (to use the expression in Rudman and Glick's (1999) article on the topic). Women who violate that norm by engaging in self promotion face the potential of backlash, which can entail psychological and material costs (as when a woman is bypassed for promotion because she is seen as "inappropriately assertive"). Thus, even a woman who feels no particular disinclination for self promotion might find it in her self interest to adopt the expected "kinder, gentler" norm (Bowles et al., 2006).[45]

Standard agency models, discussed in Sections 2 and 3, do have implications for gender in labor markets.[46] The recent work by Babcock, Niederlie, Vesterlund, and their co-authors, discussed above, adds a new and promising perspective for understanding the role of gender in organizations and labor markets. Babcock and Laschever (2003) provide extensive evidence that women's reluctance to ask often includes an unwillingness to negotiate their own salaries. It follows logically that in labor markets in which

[45] The point here is that individuals are not passive carriers of their social identities and, as Akerlof and Kranton note, there are many instances in which identities are supported by sometimes severe social sanctions meted out to those whose behavior deviates from proscribed behaviors.

[46] For example, as Bulow and Summers (1986) note, if women generally have lower labor market attachment than men (perhaps because they are more likely to withdraw from the market for bearing and raising children or elder care), efficiency wages will be less effective in motivating women than in motivating men. This leads to an equilibrium in which a higher proportion of women than men will end up in the "secondary sector." As a second example, long-hours work norms that emerge in rat race models, such as those of Landers et al. (1996) might be particularly disadvantageous to women. See Landers et al. (1997) for a discussion of this latter issue.

there is rent sharing, this psychological phenomenon contributes to male-female wage and income gaps. On the other hand, the title to Niederle and Vesterlund's (2007) paper—"Do Women Shy Away from Competition? Do Men Compete Too Much?"—suggests an important point: Cooperation, and the willingness to work hard on the behalf of others, are valuable traits, which should receive positive value in the labor markets. An important research agenda going forward is the incorporation of new findings on gender from psychology and behavioral economics into models of organizations and labor market equilibrium for the purpose of investigating these very issues.[47]

There are certainly other important identity categories that deserve attention from behavioral economists who study organizations and labor markets. *Ethnicity* and *sexual orientation* are additional identity categories that are important in many contexts, including, quite possibly, the labor market.[48] Berman's (2000) economic analysis of ultra-orthodox Jewish groups indicates a strong behavioral impact of *religious identity,* which induces many Israeli ultra-orthodox men to engage in fulltime yeshiva study into their early 40s, thereby impoverishing themselves and their families.

4.5. Miscommunication and race

Another identity category of indisputable importance is *race*. There is little theoretical work in economics that explores the role of race in organizational form and compensation practices. A very important exception is Lang's (1986) "language theory" of statistical discrimination, which focuses squarely on agency and performance within organizations. The starting point of Lang's analysis is the observation that misunderstanding and misinterpretation are common workplace problems. Lang draws on a wide body of literature in psychology and linguistics to argue that these problems

47 In this essay we focus on the effect of gender identity on agency problems, but gender identity is also likely to be very important for understanding female labor supply. In an intriguing and ingenious study, Fernandez et al. (2004), find that married women are more likely to work outside of the home if they are married to a man whose mother worked outside the home. A causal link is established by the use IV estimation, with cross-State variation in male World War II mobilization rates as the instrument.

48 In the broad social sphere, Bisin and Verdier's (2000) analysis of *ethnic* identity and intermarriage makes a strong prediction that if families value homogamous matches (matches between men and women in the same ethnic group), minority families will make greater investments in identity-preserving activities than majority families, because there is a greater chance that their children will enter heterogamous matches. In a field study of one workplace, Bandiera et al. (2009) document workplace favoritism based on nationality (presumably because of social connections between those who share language and national origin) that is costly to the firm. As for *sexual orientation,* it seems possible that when gay individuals take the (possibly very costly) break from powerful expectations to adopt a heterosexual identity and norm, this reduces costs for deviation from traditional norms along other dimensions, such as occupational choice. Along these lines, Black et al. (2000) show that during the Korean War era (1950-1954), military service rates were 12 times higher for lesbian women than other women, and Black et al. (2007) show that lesbian college graduates sort into traditionally male majors at substantially higher rates than other women. We know of no work in economics that explores implications for organizations and labor markets.

are exacerbated when managers and workers are from different "cultural" or "linguistic" groups.[49]

Following Lang's lead, Ritter and Taylor (forthcoming) consider a labor market in which there is potential for race-based workplace misunderstandings. Their focus is the possibility that this force contributes to black-white gaps in unemployment. The model of unemployment is the agency-based efficiency wage model outlined in Section 2.2.3.

Suppose that most supervisors in the US are white (perhaps because capital is disproportionately in the hands of whites in the US), and that these managers are more successful evaluating the performance of white employees than black employees. Now recall that in the efficiency wage model set out above, σ (the standard deviation of the noise) reflects the precision with which managers evaluate workers. The logic of Lang's arguments leads to the conclusion that σ is relatively higher when white managers evaluate black workers. If so, the unemployment rate will be higher for blacks than for whites.

To see the logic of the Ritter-Taylor result, recall, from (23), that in an economy with homogenous workers, an efficiency wage strategy of worker motivation leads to the following relationship between the equilibrium wage (w^E) and unemployment rate (u^E):

$$w^E = e^* + \frac{1}{\phi(z^*)}\left(\rho + \frac{F^*}{u^E}\right)\sigma. \tag{33}$$

Now suppose that black and white workers are equally productive and thus in equilibrium must be paid the same wage. Suppose also that, as discussed in the last paragraph, there is more noise in the evaluation of black workers than white workers: $\sigma^B > \sigma^W$. Then the following must hold in equilibrium:

$$e^* + \frac{1}{\phi(z^*)}\left(\rho + \frac{F^*}{u^B}\right)\sigma^B = e^* + \frac{1}{\phi(z^*)}\left(\rho + \frac{F^*}{u^W}\right)\sigma^W, \tag{34}$$

where u^B is the unemployment rate for black workers and u^W is the unemployment rate for white individuals. Clearly, $u^B > u^W$. We therefore have a potential explanation for racial differences in unemployment rates.

This model is thus consistent with evidence, such as Neal's (2006), that among men, black-white gaps in the *wage* are small when one conditions on a measurement of human capital taken when the men were youths (the AFQT), but black-white

[49] See also the excellent discussions by Cornell and Welch (1996) and Lazear (1999). The idea that minority individuals might be more difficult to assess than non-minority workers is of course also at the root of the classic work on statistical discrimination. (See, e.g., Arrow, 1998, for references to earlier literature, and a thoughtful discussion.) Austen-Smith and Fryer (2005) provide an additional important perspective.

gaps in *unemployment* are large. Ritter and Taylor (forthcoming) show that black-white unemployment gaps persist when one conditions on the AFQT, and that unemployment rates are highest for black men who attended high schools in which other students were mostly black. Under the assumption that these men are most likely to find on-the-job interactions with their boss difficult, this evidence is consistent with the model of racial differences in unemployment we have just outlined.[50]

The "miscommunication model of unemployment" outlined above calls attention to a general point that pertains broadly in models of behavioral agency: If the efficient resolution to agency problems is economically important, than labor markets will tend to reward individuals who possess scarce preferences and traits that enhance the effectiveness of firms' strategies to evaluate, monitor, and provide incentives. Bowles et al. (2001) offer a creative assessment of the labor market returns to such incentive-enhancing traits—traits that might include a low rate of time preference, an intense sense of shame at being without work, perseverance, identification with work goals, and the psychological predisposition to see personal initiative and self determination as important relative to external luck or fate (i.e., to have "internal control" rather than "external control" on Rotter's "locus of control").

Bowles et al. (2001) offer an overview of empirical evidence that wages are correlated with measures of some such traits. For example, people with a high degree of internal control earn higher wages.[51] Of course, correlation does not establish causality, and the Rotter measure may simply stand in for on-the-job productivity. Still, the authors persuasively argue that incentive-enhancing traits can matter for labor market outcomes, and may be important for understanding the large amount of unexplained variation typically observed in estimated wage regressions.

A number of recent theoretical papers in behavioral economics explore the implications of heterogeneity in agent traits along some key dimension (identification with the task, degree to which the agent is pro-social, etc.). In Section 5 below we discuss several of these papers. In each case the distribution of traits as taken to be exogenous—a reasonable approach given the goals in each paper. But, of course, in the broader scheme, many of these key traits are shaped by individuals' home and school environments. This latter point is developed in the seminal work of Bowles and Gintis (1977), who argue

[50] Grogger (2009) provides another piece of evidence consistent with the idea that impediments to black-white interactions spill over into the labor market. Even when he controls for skill and family background, blacks with speech patterns that sound distinctively black (according to anonymous listeners) are found to be relatively less successful in the labor market.

Even so, the miscommunication story we have outlined is likely a modest part of the profound racial divide in the US, as indicated by the black-white gap in unemployment and labor force participation, as well as many other economic and social dimensions.

[51] Similarly, Ritter and Taylor (forthcoming) use the Rotter measure as a control in one of their unemployment regressions, finding that men with high internal control (measured when they are young) have lower subsequent unemployment. (The same is not true of women, though.)

that if the education system is to be successful in preparing students for the labor market, the objective function ought to include the development of both cognitive skills and incentive-enhancing behaviors.

5. DUAL-PURPOSE INCENTIVES: CAN PAY DESTROY INTRINSIC MOTIVATION?

In Section 3 we discussed dual-purpose incentives as they arise in conventional models of extrinsic rewards, noting that they arise in many contexts, including the use of compensation practices to signal some otherwise-unobserved characteristic of the firm, to avoid adverse selection of workers along some otherwise-unobserved characteristic, and to deal with multi-tasking. We return in this section to the study of incentives that must do "double duty," but now the second role concerns intrinsic motives.

5.1. Pay and selection on dedication

To set the stage, we begin with an extremely simple model in which increasing pay induces adverse selection along an important dimension—dedication to the job. Our case considers a group of workers who are qualified for a particular occupation that for many is a "calling" or "vocation." There are many examples of this sort of dedication: religious ministry, policy advocacy, nursing, early childhood education, public-interest law, etc. The "calling" in this case is a potentially important form of intrinsic motivation. In this section we take a vocational inclination to be an unobserved feature of preferences. In subsequent sections, however, we allow for the possibility that the intensity of an agent's "calling" might be reinforced or eroded by the behavior of the principal or peers.

Specifically, we consider the model set up by Heyes (2005) and analyzed further by one of us (Taylor, 2007). The analysis, which uses the market for nurses as the focus of discussion, begins with simple behavioral assumptions. There are L qualified nurses, each of whom falls into two categories: (i) A proportion of these nurses, $1-\pi$, view nursing as simply a job. These individuals receive utility equal to their wage, w, and they produce value q_L on the job. (ii) The remaining proportion, π, is comprised of nurses who view work as a "vocation." They provide higher-quality nursing along an unobservable dimension, $q_H > q_L$. They also find their work fulfilling, and thus earn money metric utility m beyond the earned wage w.

Each individual has a reservation wage r which is drawn from a log concave p.d.f., $f(r)$, that has a corresponding c.d.f., $F(r)$. The function $f(\,)$ is assumed to be the same for both types of worker. Thus, at wage w, the quantity of nursing labor supplied is

$$\tilde{L}(w) = [\pi F(w+m) + (1-\pi)F(w)]L, \tag{35}$$

and the average quality is of nursing care is

$$\tilde{q}(w) = \theta q_H + (1-\theta) q_L, \tag{36}$$

where θ is the proportion of employed nurses for whom nursing is a vocation, i.e.,

$$\theta = \frac{\pi F(w+m)}{\pi F(w+m) + (1-\pi) F(w)}. \tag{37}$$

Heyes' key insight is that this latter proportion is declining in the wage.[52] Thus, the higher the wage, the lower will be the quality of services provided.

Consider an employer acting in isolation, e.g., a monopsonistic National Health Service (NHS), that hires nurses. (In different markets one could think of the Roman Catholic church setting wages for priests or nuns, or Habitat for Humanity setting wages for professional builders who take part-time positions constructing affordable housing.) Heyes shows that an employer who understands the adverse selection properties of high wages will set the wage to be lower than would otherwise be chosen. Thus, an NHS that maximizes surplus generated by nurses will operate with an apparent "shortage" of nursing, in the sense that the expected net value of product will be positive for the marginal nurse.

It is possible indeed that the principal will be driven to a corner solution, with pay set to zero. Thus, Habitat for Humanity has the well-known policy of using unpaid volunteers for many key tasks. Organization that seek to remedy injustice in the legal system often rely on *pro bono* attorney services. Historically, many religious workers take a "vow of poverty," accepting compensation at near-subsistence levels. The idea, of course, is that the lower the pay, the higher will be the dedication level of individuals willing to adopt the vocation.

Taylor (2007) extends Heyes' analysis to show that a monopsonist that seeks to maximize surplus generated by workers will always set the wage *lower* than the socially efficient level. The reason is that the monopsonist fails to take account of the surplus generated to those individuals who view their work as a vocation. Because wages are too low, too few vocationally-oriented workers end up in an occupation in which they create the greatest social value.

On the other hand, a parallel analysis in Taylor (2007) shows that if the labor market is perfectly competitive, the equilibrium wage will be inefficiently *high*. To see how this happens, notice that under the assumption that all workers must be paid the same wage,

[52] To see that point, take the derivative of θ with respect to w. The derivative has the same sign as $\frac{F(w+m)}{f(w+m)} - \frac{F(w)}{f(w)}$, which is negative for a log concave function.

a social planner would want to maximize

$$\pi\left[(q_H+m)F(w_H+m)+\int_{w_H+m}^{\bar{r}} rf(r)\mathrm{d}r\right] + (1-\pi)\left[q_L F(w_L)+\int_{w_L}^{\bar{r}} rf(r)\mathrm{d}r\right], \tag{38}$$

where r is taken to be the value of the worker in some other capacity (with $\bar{r}$ being the highest value in the distribution). Maximization of (38) leads to the wage being set to be the weighted sum,

$$w^* = \hat{\theta}q_H + (1-\hat{\theta})q_L, \quad \text{with } \hat{\theta} = \left[\frac{\pi f(w+m)}{\pi f(w+m)+(1-\pi)f(w)}\right]. \tag{39}$$

Next notice that in a competitive market the wage instead will equal average productivity,

$$w^c = \tilde{\theta}q_H + (1-\tilde{\theta})q_L, \quad \text{with } \tilde{\theta} = \left[\frac{\pi F(w+m)}{\pi F(w+m)+(1-\pi)F(w)}\right]. \tag{40}$$

Using the property of log concavity given in Footnote 52, we can compare wages in (39) and (40), finding that $w^c > w^*$. The problem with the competitive market is that each firm makes hiring decisions on the basis of *average* market productivity. A social planner would instead make decisions on the basis of the productivity of the *marginal* worker, i.e., would take account of the fact that as the wage increases in the market, the productivity of the marginal worker declines.

This simple model serves as a first illustration of an important point that reappears throughout this section of our essay: Pay policies can affect intrinsic motivation, often in surprising ways. Here, high pay *reduces* intrinsic motivation in a workforce in a particularly transparent way. Low-pay environments attract workers for whom the job is a vocation—workers who have an intrinsic inclination to provide high-quality service. The higher the pay, the greater will be the proportion of workers for whom the job is simply a job, i.e., workers who will provide lower-quality service.

A particularly interesting feature of this simple behavioral model is that markets can lead to wages being either too high or too low relative to an efficient benchmark, depending on the market's structure. To see the logic of this point, consider this question: "If you were a falsely convicted death-row inmate, would you rather be in a State in which you must rely on an organization that reviews cases using *pro bono* attorney services, or in a State that purchases legal services on the competitive market?" In the State that relies on *pro bono* services, attorneys who work on death-penalty cases will be highly dedicated to justice, and will provide excellent legal aid, but that aid will

be in short supply. In contrast, in a State that purchases legal services for death-row inmates, access to attorneys may be more extensive, but those attorneys will have a lower expected level of dedication. Our model shows that in the State that relies on *pro bono* attorneys, the wage is too low and the quality level too high relative to the efficient benchmark. But in the State that uses the competitive market, the wage is too high and the quality too low. Theory alone does not identify the socially preferred second-best outcome.

Recently, a number of papers have examined models in which agents differ in their level of intrinsic motivation. Delfgaauw and Dur (2007), for example, have a wage posting model in which a monopsonist faces the same tension discussed above: the higher the posted wage, the higher the probability of filling a vacancy, but the lower the expected motivation level of workers who apply. In their model, workers have private information about their utility—information which they may wish to signal to or conceal from an employer. Besley and Ghatak (2005) and Delfgaauw and Dur (2008) study public sector employment under the assumption that some agents have a "public service motivation" that takes the form of intrinsic value derived from making a contribution to one's organization.[53]

The model we have examined in this section omits, obviously, several relevant issues that merit further consideration. For instance, the set-up abstracts from the core problem of agency; workers are simply assumed to supply effort on the basis of their internal intrinsic values. Second, workers are assumed to be *steadfast*; motivation is not affected by the actions of those around them. Thus, a worker who is inclined to provide high quality service is not de-motivated when she is surrounded by others who provide low quality service. In short, the analysis abstracts from "social preferences" of the sort discussed in Section 4. Third, the model does not take account of the possibility that a worker's motivation can be affected by attributions the agent might place on the intentions of the principal. It is easy to see that such attributions might be germane, though, in the case of the "service motivations" we have been discussing.[54] Finally, the set-up does not allow for the possibility that a worker's intrinsic motivation (m in the model above) can be reinforced or undermined by pay policies. We turn next to models that take up such issues.

[53] See also Prendergast (2007), who sets out a model in which there is variation in the degree to which agents care about the outcome of some action they might take, as when bureaucrats vary in the extent to which they have altruism and empathy for individuals they are serving. One example he develops concerns social workers hired to determine eligibility for public assistance programs. While a "client-serving ethic" is important for this occupation generally, that same trait may be an impediment for the bureaucratic task at hand.

[54] For example, when an organization hires motivated agents to pursue some jointly-shared social goal, agents must believe that they indeed are advancing that goal. Presumably, religiously-oriented individuals will be demoralized if they discover that are working for a corrupt church. See, e.g., Besley and Ghatak (2005) for a discussion along these lines.

5.2. Social preferences, conformism, and the principal's use of extrinsic rewards

Recent work by Sliwka (2007) considers the question of agency in a model that draws on a "social preference framework," i.e., allows for agents' motivation to be shaped in part on the behavior of those around them. As in the model set out in the last section, there is heterogeneity in worker motivation, and an agent's motivational inclinations are initially hidden to the principal and to other agents. There are two types of "steadfast" agents. One type is "strictly selfish;" these are agents who care only about their own payoff. Agents of the second type are "fair" in the sense that they care about the wellbeing of others; specifically, for these agents, utility is increasing in the principal's payoff. The key innovation is to assume that there is yet a third group, "conformists," whose inclination toward fairness depends the values of those around them. To keep matters simple, Sliwka assumes that when a conformist learns what agent type is in the majority, the conformist behaves like the majority-type agent.

In this set-up, a principal who understands that most of his steadfast workers are fair, might be able to use compensation policies as a credible signal to "conformists." In turn, if conformists believe that others around them are "fair" they behave like fair agents.

To see how this works, we set up a simple example similar to that developed by Sliwka. In our example, the principal first posts a policy that specifies wage as a function of effort (which is assumed to be observable *ex post*), $w(e)$. An agent's best response to the announced policy depends, of course, on his preferences over effort and money, and those preferences in turn vary by type. In particular,

$$\text{utility} = \begin{cases} w(e) - \dfrac{e^2}{2} & \text{for a steadfast selfish agent, and} \\ w(e) - \dfrac{e^2}{2} + \mu\pi & \text{for a steadfast fair agent,} \end{cases} \tag{41}$$

where μ reflects a fair agent's level of identification with the principal's objective (with $0 < \mu < 1$).

We assume that the principal sets compensation to have a fixed component and a "bonus" that is a linear function of effort, $w(e) = w_0 + \beta e$. Given that types are unobserved, the principal's posted wage-bonus policy applies to all agents. It is easy to see that for an announced compensation policy, best responses are

$$\hat{e}(\beta) = \begin{cases} \beta & \text{for a steadfast selfish agent, and} \\ (1-\mu)\beta + \mu & \text{for a steadfast fair agent.} \end{cases} \tag{42}$$

With this in mind, consider a profit-maximizing principal who earns surplus

$$\pi = e - w(e) \tag{43}$$

for a given agent. Given the best responses in (42), it is clear that effort is increasing in the incentive intensity; $\hat{e}'(\beta) > 0$ for both types of steadfast agent. Sufficiently high effort-contingent bonuses would seem to be in order.[55] Remarkably, it might nonetheless be in the principal's best interest instead to set β equal to 0 and increase the baseline wage, i.e., to rely solely on low-powered incentives.

The key is the emergence of a separating equilibrium in which conformists become convinced that most steadfast agents are *fair.* It is assumed that the firm has private information about the proportion of steadfast agents who belong to each type, which for simplicity is taken to have a *low* value or a *high* value. It is a matter of simple algebra to confirm that there are parameter values for which the following holds: If the firm has a *low* number of fair agents, it pays a high bonus, $\beta > 0$, and a low wage. If the firm has a *high* number of fair agents, it pays no bonus, $\beta = 0$, and a relatively higher wage, i.e., it uses low-powered incentives. Here low-powered incentives serve as a credible signal, so conformists follow suit and behave like fair agents. This makes sense, because a principal who has a *high* fraction of steadfast fair agents will incur a smaller loss than a principal with a *low* fraction of fair agents when it sets the bonus to 0. A willingness by the principal to raise the fixed wage further strengthens the signal. Conformists, in response to this credible signal, behave like fair agents. Profit for this firm is higher than if it used higher-powered incentives.

The operating logic of the model is like the Taylor-Ritter model discussed in Section 3.1, in which the firm uses compensation policy to signal hidden information about itself (i.e., the firm's financial fitness). In that model, low-powered incentives signal relatively good financial fitness, which allows the firm increased profitability.[56] Here, the principal's hidden information is the mix of worker type. There is a cost to low-powered incentives, of course, as effort is set by all agents to be less than first-best. But the low-powered incentive persuades some workers—the "conformists"—to behave in an altruistic fashion, when they would otherwise have not.

Sliwka interprets his model as generating "trust as a signal of a social norm." In Sliwka's setting, the principal observes effort and can, if he chooses, condition rewards on effort. By setting no explicit incentives, the principal expresses trust in his workers. This trust directs a social norm. Some workers are more generous in their efforts than they would have been if they perceived a different norm.

The most intriguing possibility in the Sliwka's model is that *monetary incentives crowd out intrinsic motivation*. If a firm moves from a "high trust" low-powered incentive

[55] With perfect information, the bonus would be $\beta_S = \frac{1}{2}$ for a selfish agent and $\beta_F = \frac{1}{2} - \frac{\mu}{2(1-\mu)} < \beta_S$ for a fair agent. Ideally, the principal prefers a larger bonus for a selfish agent, but will want to have a positive bonus for fair agents as well as long as $\mu < \frac{1}{2}$. Notice also that fair agents have "intrinsic motives;" they provide effort even when the bonus is 0.

[56] Similar logic pertains in Spier's (1992) model, in which a principal knows more about the profitability and riskiness of a project than does an agent, and in Allen and Gale (1992), in which a supplier has superior information about his ability to distort a signal of production costs.

scheme to a "low trust" high-powered incentive scheme, the firm shifts the norm and undermines the intrinsic portion of worker's motivation (the "social component" in the utility of a worker who would otherwise behave as a "fair agent"). Sliwka develops his theory further by looking at employee self-selection into firms. Here again, low-powered incentives can serve to attract workers with high intrinsic motives (fair agents), which serves to reinforce the positive work norms that influence those who conform to others.

The key behavioral underpinning of the Sliwka model is the observation that many people seem to want to conform to those around them. As we have noted, there is considerable evidence about the key component of this story—that many people are influenced by *norms*. For example, some individuals feel bad about a particular action only in situations in which they think others would experience remorse for that same action.[57] In previous sections of this paper, we cited empirical work supportive of the social forces that create norms, e.g., studies by Ichino and Maggi (2000), in which worker absenteeism in an Italian bank was affected by the absenteeism of those around them, and Mas and Moretti (2009), in which effort by supermarket checkout workers was affected by other similar workers in their sightline. Yet another study, by Bandiera et al. (2009), shows that the productivity of farm workers is affected by the productivity of friends on the job. Jackson and Bruegmann (2009) document peer learning for teachers, which might be read as providing additional evidence on conformity to norms. Of course, considerably more empirical work will be required to know if conformism plays a sufficiently strong role to generate in real-world organizations the crowding out of intrinsic motivation predicted by the Sliwka model.

5.3. Extrinsic incentives when agents value the principal's esteem

Ellingsen and Johannesson (2008) present a model that, like Sliwka's, relies on social preferences. Also, like Sliwka's, their model opens up the possibility that extrinsic incentives can undermine valuable intrinsic motivations.

The key innovation in the Ellingsen-Johannesson model is the postulate that human motivation is often rooted in *social esteem*—the desire to be well regarded by others. In this conception, an agent reasons as follows: "I wish for others to hold a high opinion of me. While I cannot know with certainty what others think of me, I do have beliefs about what others think, and these beliefs about others' opinions are an important source of pleasure or discomfort." The identity of the audience that the agent wishes to impress plays a key role in this model, and the agent might well have multiple audiences. For

[57] There is a great deal of empirical work across disciplines on norms. One particularly evocative story is told by Fisman and Miguel (2007): When United Nations diplomats in New York were given immunity for parking violations, violations were much higher among diplomats from countries that have high levels of corruption than from countries that have low levels of corruption. This suggests a powerful role for cultural norms. At a theoretical level, work by Bernheim (1994) is important. Fischer and Huddart (2008) discusses the role of endogenous social norms on organizational design.

example, a college professor might care about opinions of her students, her dean, other professors in her department, and/or colleagues in the profession more generally. She desires the *respect* of the intended audience(s), i.e., gains utility if she believes that others think highly of her.

Ellingsen and Johannesson focus on the case in which the relevant audience to an agent is the principal.[58] To simplify use of personal pronouns, we let the principal be male and the agent be female. In the model, then, the agent's utility depends on the "respect" she earns from the principal, which is defined to be *her beliefs about his beliefs about her.*

A simple example shows how the desire to earn respect can affect an agent's effort decisions.[59] Suppose there are two types of agent, "talented" and "untalented," and type is not initially observable to the principal. An agent hired by a principal is paid an agreed-upon wage w, and then chooses any effort level she likes, $e \geq 0$. We suppose that her utility is the sum of three components: (1) compensation w, (2) the cost of effort, which is $-c_1 e$ for a talented worker and $-c_2 e$ for an untalented worker, with $c_2 > c_1 > 0$, and (3) "respect" of the principal, which has value $rp(e)$, where r is a positive constant and $p(e)$ is the agent's subjective probabilistic assessment of the principal's belief that the agent is talented. In sum,

$$\text{utility} = \begin{cases} w - c_1 e + rp(e) & \text{for a talented agent, and} \\ w - c_2 e + rp(e) & \text{for an untalented agent.} \end{cases} \tag{44}$$

Now, given (44), an *untalented* agent is clearly better off supplying effort 0 than supplying effort greater than $\frac{r}{c_2}$, even if the higher effort level would "earn maximum respect" (i.e., would induce the principal to believe with probability 1 that the agent is talented). So we have a separating equilibrium satisfying the Intuitive Criterion if a *talented* agent supplies effort $\frac{r}{c_2}$, which is just high enough so that the *untalented* agent will decline to mimic.[60]

In this example, a talented agent provides positive effort. She is *not* motivated by her own material wellbeing, as she would be in a standard principal agent model. Nor is she motivated by an innate desire to see the principal's wellbeing improve, as with the other-regarding preferences assumed in the Sliwka model (or other such models discussed in Section 4). She provides effort because by so doing she can be confident that the principal holds a high opinion of her. Put another way, she is motivated by *social esteem*—the desire to earn respect.

58 Kandel and Lazear's (1992) important work on peer pressure, in contrast, focuses on the case in which an agent values the regard of co-workers.

59 The example set out here follows Ellingsen and Johannesson (2007).

60 For such an equilibrium to exist, we need for the difference between c_2 and c_1 to be large enough to support separate actions by the two types. Suppose it is common knowledge that proportion π of agents are talented. If all agents were to supply $e = 0$, agents would earn respect $r\pi$. But in such a proposed equilibrium, it would be worth it to a maverick talented agent to play $e = \frac{r}{c_2}$, and thereby earn respect r, only if $r - c_1 \frac{r}{c_2} > \pi r$, which boils down to $(1-\pi)c_2 > c_1$.

With this basic logic in place, we can set out the Ellingsen-Johannesson model of principal-agent interaction. The model is built around three components. First, agents and principals hold *social preferences.* They care about their own material wellbeing as well as well as the material wellbeing of others. Second, there is unobserved heterogeneity in the extent to which agents and principals value others' wellbeing. In particular there are two types on people, who vary in the extent to which they are pro-social. Third, and most distinctively, both the agent and the principal are motivated by *social esteem,* so the agent cares about what the principal thinks about her, and the principal cares about what the agent thinks of him. Both want to be thought of as pro-social by the other. Moreover, the agent's concern about the principal's opinion is highest if she thinks highly of him, i.e., believes it is likely that he is highly pro-social. Similarly, the principal places greater weight on the agent's opinion if he believes that she is highly pro-social.

With these assumptions in place, Ellingsen and Johannesson examine the equilibrium of a game in which the principal takes an initial action (e.g., makes a wage offer, or makes a decision about how much discretion to allow the agent in her work), and then the agent takes an action which affects both her material wellbeing and the principal's wellbeing. As in the simpler example in the preceding paragraphs, there is a set of parameters on preferences and the distribution of types such that a separating equilibrium emerges that satisfies the Intuitive Criterion. In that equilibrium a pro-social principal can take a credible action that signals that he is pro-social, and, if she is sufficiently pro-social, the agent responds with an action that benefits the principal. The key driving behavioral force is that a pro-social agent wishes to be highly regarded by a pro-social principal. Having learned that the principal is pro-social, the agent takes a pro-social action herself as a means of securing the knowledge that the principal believes her to indeed be a pro-social individual.

Figure 2 provides a nice illustration of the behavior predicted in this model. The game presented is the two player sequential "trust game" of McCabe et al. (2003):

In Case 1, the first player (the principal) can choose "not trust" (NT), which leads to payoffs (20, 20) for the principal and agent respectively, or "trust" (T), which accords discretion to the agent. If the principal plays T, the agent can reward the trust (R), giving payoffs (25, 25) or not (N), giving payoffs (15, 30). With conventional preferences, the subgame perfect equilibrium is clearly "not trust." However, in the McCabe-Rigdon-Smith experiments, many principals chose T, and most agents responded by rewarding such trust by playing R.

In Case 2, the principal has no choice but to play T. In contrast to Case 1, here most agents responded by playing N. Thus, the intentionality of the principal's trust appears to matter for the agent's response.

The Ellingsen-Johannesson model provides a rationale for these observed outcomes in the trust game. Start with Case 1. There are parameters in the model such that two key conditions are met. First, the principal will play T only if he is sufficiently pro-social,

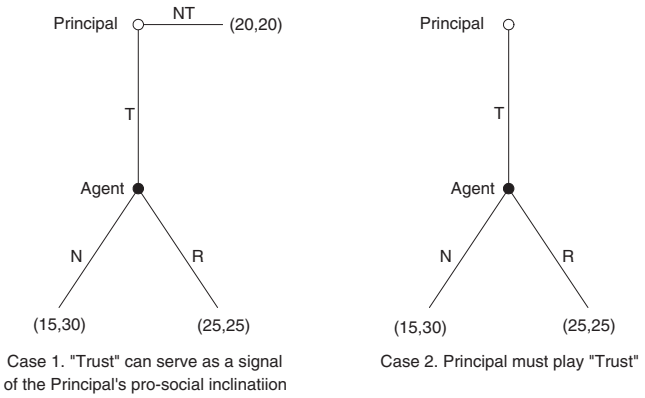


Figure 2 ***A principal agent game of trust.***

i.e., only if he cares sufficiently about the wellbeing of the agent. Second, a pro-social agent, having received a credible signal that the principal is pro-social, cares sufficiently about the respect of the principal that she in turn takes the action R, confirming that she is pro-social.

In contrast, in Case 2 the principal has no opportunity to signal that he is pro-social. In turn, the agent cares less about his respect, and so she plays N.

Ellingsen and Johannesson (2008) show, using similar logic, that their model predicts behavior consistent with Falk and Kosfeld's (2006) experimental evidence on the hidden "cost of control" in a principal agent game. In the Falk-Kosfeld game, agents are given an endowment of 120 and can transfer $x \leq 120$ to the principal, who in turn receives payoff $2x$, thus resulting in payoffs $(2x, 120 - x)$. The important twist is that in some conditions of the game, there is a first stage in which the principal can play "control" by imposing a minimum transfer (e.g., a transfer of 10) from the agent to the principal, or can choose instead to "trust." Agents motivated solely by material gain would always play the minimum available x, and knowing this, the principal would always "control" to the maximum extent allowed. But, in fact, consistent with the Ellingsen-Johannesson set-up, many principals sent a signal of being pro-social themselves by choosing "trust" when they are allowed to do so, and in such cases many agents responded with larger values of x than if the principal had played "control."

One nice way to see the distinctive contribution of the "esteem model" is to view it in the context of Akerlof's (1982) concept of gift exchange. Recall that in Akerlof's model (discussed above in Section 4.2), an agent's best response to a sufficiently generous "gift" by the principal is to reciprocate by providing high effort. The agent's motivation to do so is captured in a clear, but stripped down fashion—with a utility function in which the agent experiences disutility only when her effort level exceeds a psychologically determined threshold (the effort "norm"). Ellingsen and Johnnesson take an additional

step, positing an explicitly specified behavioral mechanism ("esteem") that drives this motivation. This approach, the authors show, allows them to predict gift exchange behavior. But there are two advantages to the Ellingsen-Johannesson model:

First, the model gives a clear way of understanding the role of a principal's *intentions* in shaping agents' responses. This is important, given experimental evidence (such as (Charness, 2004)) that intentionality is important to understanding gift exchange.

Second, the model provides a rigorous way of approaching an important and under-appreciated aspect of principal agent problems as they apply in the workplace—the delegation of decision rights. Essentially, the delegation of consequential actions to agents plays the role of a "gift" here, and provides the agent with the opportunity to earn the respect of the principal. In contrast, highly intrusive job design diminishes intrinsic motivation.

These ideas are potentially valuable for understanding otherwise-inexplicable practices within organizations. For example, charitable organizations like Habitat for Humanity often rely on volunteers who receive little or no pay, and then delegate key decisions to these same individuals. By providing low pay, we have suggested (in Section 5.1 above), the organization is less likely to attract opportunists. Then the "esteem model" shows the important advantages to relinquishing bureaucratic control. A second example is the widespread use of "psychological contracts" (Rousseau, 1995) in which contracting parties find it advantageous to leave many elements unspecified, relying instead on mutual goodwill.

Conversely, the esteem model indicates why high pay and clearly delineated direction might be required in other circumstances. This can happen when the desire for esteem leads agents to have intrinsic motivation that works at cross purposes to the principal. For example, very strong financial incentives might be required to induce a physician to cut costs if the physician values the esteem of her patients more than the esteem of her boss (e.g., the managed care organization she works for).[61]

5.4. Extrinsic rewards and reputation

The Ellingsen-Johannesson model we have just discussed is one of a number of recent papers that focus on the interaction of pro-social motivation and reputation or esteem. This literature starts with the observation that people undertake altruistic and reciprocal actions (in the workplace and elsewhere), but recognizes that such behavior is often difficult to rationalize solely by other-regrading preferences. The degree to which

[61] Quite possibly, carefully constructed models of esteem formation can make predictions about the interaction between organizational design and society's class structure. Esteem motives, after all, break down if the principal is not a member of the audience, i.e., the class of individuals whose respect the agent values. Thus, if female identity is formed in a way that leads women to especially value the esteem of men, male bosses would have a distinct advantage over female bosses when supervising women. Bosses in high positions in racial hierarchies, ethnic hierarchies, or other socially determined hierarchies would be similarly advantaged.

people undertake pro-social behavior often depends on the social context and economic environment, and is driven in part by the desire to be highly regarded by others.

An important contribution to this literature is the recent work of Bénabou and Tirole (2006). Among the remarkable insights of this paper is a clear demonstration that extrinsic rewards can undermine intrinsic motivation when people care about reputation.

Consider the following anecdote: One of us has a particularly personable colleague who was asked by the dean to accept a somewhat onerous task—advising masters students—that would have high value to his colleagues. In exchange, the dean offered a $2000 bonus. The professor replied that it was certainly not worth taking on the task for $2000, but that he would be willing to do the job for free!

To demonstrate how the Bénabou and Tirole approach explains the behavior of this public-spirited professor, we set up a simplified case of their more general model. We suppose that a principal asks agents to undertake a pro-social activity by providing effort $e = 1$. The agent can decline, instead providing $e = 0$. The agent's effort choice is observable by all, including members of an audience whose opinion matters to the agent.[62]

The agent's effort decision is assumed to affect his utility via four channels, which for simplicity are taken to be additive: (1) The agent is other-regarding, and so earns direct utility, $v_e e$, from providing effort e; (2) he stands to earn a material reward in the form of a bonus of $b \geq 0$, which provides utility $v_m be$ (where v_m is the marginal utility of money); (3) he faces an effort cost of ce; and, most distinctively, (4) he stands to gain from the reputation-enhancing effect of his effort choice.

Central to the Bénabou-Tirole model are the following two assumptions, which drive development of "reputation." First, people differ in the extent to which they have other-regarding inclinations, and in the extent to which they value money. So each individual's set of preference parameters, v_e and v_m, is drawn from a known distribution.[63] An agent's preference type can be thought of as his or her "identity." Second, reputation is taken to be other's views of one's own identity. This reputation is *increasing* in the degree to which one is seen as having concern for others (having a high value of v_e) and decreasing in the degree to which he is seen as materialistic (having a high value of v_m).[64] Thus, reputation is taken to be $R(e, b) = \mu_e E[v_e|e, b] - \mu_w E[v_m|e, b]$, where μ_e and μ_w are weights that reflect the degree of image-consciousness (and are taken to be common knowledge constants here).

62 To simplify matters, suppose that the agent is already in the principal's employ and is now being asked to undertake a task that was not originally part of the job (as in the example of the public-spirited professor). It would be a worthwhile task to apply the Bénabou-Tirole model in a labor market generally (which would require attention to participation constraints, and to the way in which an announced compensation policy might affect selection into a firm).

63 To keep things simple here, we suppose that the parameters are independent.

64 Concern for others ("kindness") and moderation in materialistic pursuit ("temperance") are but two of the seven virtues. By leaving the other five virtues unstudied, perhaps Bénabou and Tirole signal "patience" (leaving them for future work) and "humility" (deference to other behavioral economists who wish to study those virtues).

To summarize,

$$\text{utility} = \begin{cases} v_e + v_m b - c + R(1, b) & \text{if } e = 1 \text{ and} \\ R(0, b) & \text{if } e = 0. \end{cases} \tag{45}$$

So our agent provides effort if

$$v_e + v_m b + [R(1, b) - R(0, b)] > c. \tag{46}$$

The three terms on the left-hand side of (46) are, respectively, the agent's intrinsic, extrinsic, and reputational motivations. Effort is provided when the sum of these motivations exceeds the cost of providing effort.

What makes matters interesting here is that the reputation the agent earns depends on the level of the bonus b chosen by the principal. To see this point, consider first the effort decision if the principal chooses $b = 0$, so that motivation is strictly intrinsic and reputational. Then the agent provides effort only if his "concern for others" exceeds the cut-off $\hat{v}_e$, where

$$\hat{v}_e \equiv c - [R(1, 0) - R(0, 0)]. \tag{47}$$

Such highly pro-social identity types lie to the right of the vertical "No Bonus" line in Fig. 3. Notice that in this situation the agent's audience can draw an informative inference about the agent's v_e (i.e., can infer if the agent is in an other-regarding pool to the right of $\hat{v}_e$ or narcissistic pool to the left of $\hat{v}_e$) simply by observing the effort level, but can learn nothing about his materialism v_m.

Now suppose that the principal provides a bonus b. For the moment, ignore any impact on reputation. The bonus has no effect on an agent with $v_m = 0$, of course, but for all other agents (those with $v_m > 0$) the bonus is motivating. Agents now provide effort if they have an identity that lies to the right of the negatively sloped line marked "Positive Bonus ($b > 0$), Unadjusted for Reputation." Absent reputational effects, an extrinsic reward expands the pool of agents providing effort.

However, this is not the end of the story. Inspection of Fig. 3 makes it is clear that the average level of v_e (concern for others) has declined in the pool of agents providing effort and also in the pool of agents not providing effort. It is similarly clear that the average level of v_m (greed) has risen for the pool of those providing effort and declined for the pool not providing effort. So the overall effect on reputation is ambiguous. The most interesting possibility is that the effect of the bonus is to drag down overall reputation for those providing effort. This is demotivating. In Fig. 3 this is illustrated by the parallel shift to the right in the sloped line dividing those who provide effort and those who do not, with $\tilde{v}_e$ now giving the value of v_e that separates the two pools for individuals along

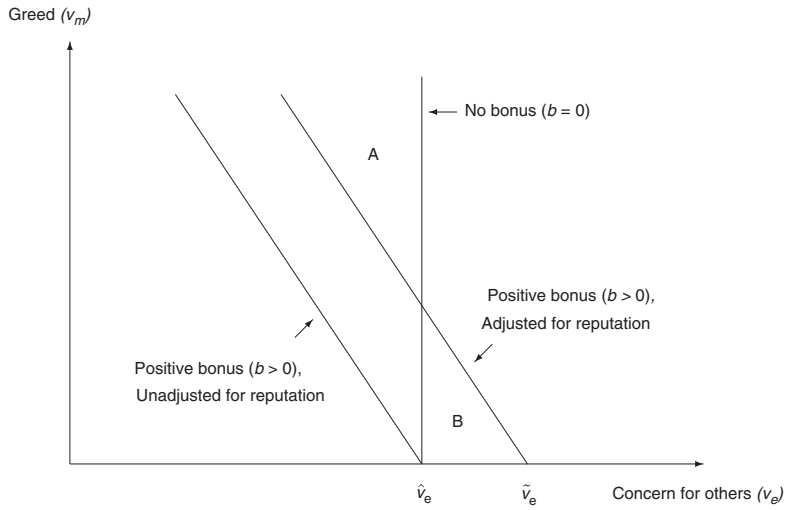


Figure 3 ***The effects of an extrinsic reward on the pool of agents providing effort.***

the horizontal axis (i.e., for individuals with $v_m = 0$). On net, effects of the bonus are two-fold: Some identity types—those in Area A of Fig. 3—switch behavior to providing effort. Others—those in Area B—are induced to switch from effort provision to *not* providing effort. Overall, an extrinsic reward can increase or decrease effort, depending on the distribution of identity types.

The public-spirited professor in our anecdote is apparently an individual with an identity of the sort represented by Area B. Such a person has a relatively high concern for others and a relatively low level of greed. By offering a bonus for the task, the dean deprived the professor of the opportunity to demonstrate his public spiritedness.

In short, in the Bénabou-Tirole model, extrinsic rewards can spoil the reputational value pro-social action, thereby crowding out intrinsic motivation. The logic of the model is one of "signal extraction." People take pro-social actions in part to signal one's own identity to others. Extrinsic rewards, even very small extrinsic rewards, can serve to increase the noise-to-signal ratio of such actions.[65]

One of the most interesting ideas in the Bénabou and Tirole model appears when the authors reinterpret the "reputational" terms in (46) to instead be the reinforcement of

[65] Indeed, in more general versions of their model, Bénabou and Tirole show that extrinsic rewards can reverse the sign of the signal! Armed with this logic, the authors establish interesting and surprising insights around the use of such non-monetary motivators as *praise* and *shame*. They show, for example, that the excessive use of praise can backfire if pro-social behavior "becomes suspected of being motivated by appearances." They also study the equilibrium development of social norms.

one's own *self image*. The idea is described as follows:

> *When making a decisions affecting others' welfare, an individual will often engage in a self-assessment: "How important is it for me to contribute to the public good? How much do I care about money? What are my real values?" Later on, however, this information may no longer be perfectly "accessible" in memory—in fact, there will often be strong incentives to recall it in a self-serving way. Actions, by contrast, are much easier to remember than their underlying motives, making it rational to define oneself partly through one's past choices: "I am the kind of person who behaves in this way" (Bénabou and Tirole, 2006).*

Thus, the public-spirited professor might have found the offer of a $2000 bonus to be demotivating even if his colleagues were unaware of the bonus. Accepting the task without pay, in this conception, served to reinforce his identity; the professor can look at himself in the mirror and argue convincingly, "I must be a pro-social person. Otherwise I wouldn't have taken on this task with no pay."[66]

The striking prediction that extrinsic rewards can crowd out desired behaviors does have empirical support. One widely cited example in economics is work by Gneezy and Rustichini (2000b) showing that the imposition of a monetary penalty for late child pick-up at a daycare center increased the likelihood of late pick-up. Bénabou and Tirole's reputation/self-respect model strikes us as applicable here. When there was no explicit monetary penalty for on-time pick-up, parents were presumably motivated by genuine concern for daycare center workers and by a desire to project to others (or to oneself) character traits of *responsibility* or *concern for others*. The imposition of a monetary penalty of course increased the inclination for on-time pick-up among those parents with a materialistic orientation and a low level of concern for others. This very fact led parents more generally to no longer view on-time pick-up as a reliable signal of kind and responsible identity, and so reduced the strength of those motivating forces. Similar arguments apply to Gneezy and Rustichini's (2000a) demonstration that extrinsic incentives reduced effort by school children collecting donations for a charitable organization.[67]

Evidence of a potentially important form of crowding out is also found also in Frey and Oberholzer-Gee's (1997) analysis of public reaction to the siting of a nuclear waste facility in one's community. Their paper indicates that the provision of substantial compensation to residents of a host community reduces willingness to accept such a facility. The reputation/self-respect model might speak to these results, but it is quite possible that mechanisms described in Bénabou and Tirole (2003) are more germane.

66 Bénabou and Tirole (2006) note that the key idea—"that individuals take their actions as diagnostic of their preferences"—is found in psychology in Bem's (1972) *self-perception theory* and is related to the Festinger and Carlsmith's (1959) theory of *cognitive dissonance*.

67 Frey and Jegen (2001) provide further reference to the literature, and discuss crowding effects from an economic perspective. See also Frey (1997).

In that paper the authors set up a problem in which a principal seeks to motivate an agent to take a desired action in an environment in which the principal has better information than does the agent about some crucial aspect of the task—for example, the cost the agent will incur if she undertakes the task, the personal satisfaction she will experience if the task is completed successfully, or the likelihood that the agent will indeed be successful at the task. In this setting, the offer of a substantial monetary reward for some action can signal "bad news" to the agent about one of the elements of the action. Thus, if residents of a potential host community are asked to site a nuclear waste facility *and* are offered substantial compensation for doing so, that compensation might be seen as "bad news" about the eventual consequences of the facility.

The two papers we highlight in this section of our paper, Bénabou and Tirole (2003, 2006), are but two of a number of recent contributions in behavioral economics that might form solid building blocks for a new generation of behavioral principal agency models.[68] The challenge going forward is to place the psychological subtleties introduced in these new economic models into workable (and testable) theories of firm organization and labor markets. It is important to have carefully constructed, psychologically correct models in behavioral economics, but important also to work forward to understand the implications of these models for the allocation of resources in markets and in society broadly.[69]

5.5. A concluding puzzle

The economic approach to agency places a primary emphasis on the use of material incentives (pay, promotion, etc.) as devices to resolve principal agent problems. The economic literature offers a rich and varied set of evidence in support of the critical efficacy and importance of material incentives. The theoretical literature reinforces these empirical findings. There are many situations in which firms eschew high-powered incentives, but for the most part this is the result of incentives having a powerful effect on behaviors. It is possible, as we have seen, to construct models where extrinsic rewards undermine intrinsic motives, but these models appear largely as elaborations and qualifications of the fundamental message: well designed extrinsic rewards are crucial to the resolution of fundamental and ubiquitous agency problems.

68 Among the many other potentially relevant examples are the models of social image in Bernheim and Severinov (2003) and Andreoni and Bernheim (2009). The Bernheim-Severinov model is designed to explain the common practice of equal division of bequests. The model posits that children care about the extent to which they are loved relative to other siblings, and then studies bequests as a mechanism by which parents can signal love. Equilibrium behavior tends to pool at equal bequest division. Similar logic might explain the frequent organizational practice of equality in treatment (pay, work conditions, etc.) of workers who might differ quite widely in productivity. Andreoni and Bernheim's refinement of these ideas might serve as a valuable microfoundation for studying the role of fairness (e.g., Fehr and Schmidt, 1999) in principal agent relationships.

69 The beautiful work of Akerlof and his co-authors—in papers on the economic implications of reciprocal motives, cognitive dissonance, social distance and identity—provides a template in this regard.

Things are quite different in the field of psychology. Here there has accumulated a vast amount of evidence that extrinsic rewards actually undermine intrinsic motives.[70] What explains the difference?

One important part of the explanation is a cross-disciplinary difference in the definition of *intrinsic motives*. Psychologists typically view as extrinsic any sort of action undertaken for instrumental reasons. Thus many of the pro-social and other regarding preferences we discuss in Sections 4 and 5 would be regarded as part of an extrinsic reward system in the psychology literature.[71]

Economics is concerned with the efficient use of society's material resources. In societies characterized by specialization and a sophisticated division of labor, almost all economic activity involves some degree of instrumental motives. Thus by defining the notion of intrinsic rewards so narrowly, psychologist's have restricted their attention to a very small subset of economically relevant behaviors.

The focus, as peculiar is it might appear from an economist's perspective, makes perfectly good sense from the perspective of psychology. After all, psychology is concerned with understanding the reward structures that drive human behavior. Why then should psychology privilege economically relevant motives?

There is another important difference in the ways that psychology and economics analyze extrinsic and intrinsic motivation: the handling of *autonomy*. Deci, Ryan, and other psychologists argue that feelings of autonomy and competence are fundamental to human happiness. To the extent that they cause people to become accustomed to responding to rewards rather than their own intrinsic drive for self-realization, extrinsic rewards undermine a fundamental determinant of psychological wellbeing.[72]

70 Important theoretical constructs include Lepper et al.'s (1973) *overjustification theory* and Deci and Ryan's (1985) *self-determination theory*. A large number of carefully constructed experiments provide evidence favoring these theories, including many that demonstrate crowding out of intrinsic motivation (e.g., Deci et al., 1999).

71 For instance, in Ryan and Deci's (2000) taxonomy, *intrinsic motivation* is reserved for "the doing of an activity for its inherent satisfactions rather than for some separable consequences." *Extrinsic motivation,* on the other hand, "pertains whenever an activity is done in order to attain some separable outcome." Such extrinsic motivation includes "external regulation" with a material reward or punishment, but also includes "introjection," which focuses on approval from others or from oneself, and also, remarkably, "integrated regulation," which occurs when an agent comes to assimilate the external driver as an internal driver. To quote Ryan and Deci (2000), "The more one internalizes the reasons for an action and assimilates them to the self, the more one's extrinsically motivated actions become self-determined. Integrated forms of motivation share many qualities with intrinsic motivation, being both autonomous and unconflicted. However, they are still extrinsic because behavior motivated by integrated regulation is done for its presumed instrumental value with respect to some outcome that is separate from the behavior, even though it is volitional and valued by the self."

72 Deci and Ryan's theory of self-determination, for example, emphasizes the innate psychological needs for a sense of *competence* and *autonomy*. The authors suggest that "interpersonal events and structures (e.g., rewards, communications, feedback) that conduce toward *feelings of competence* during action can enhance intrinsic motivation for that action because they allow satisfaction of the basic psychological need for competence. Accordingly, for example, optimal challenges, effectance promoting feedback, and freedom from demeaning evaluations are all predicted to facilitate intrinsic motivation" (Ryan and Deci, 2000).

Positive economics, in contrast, conceives of autonomy simply as a means for achieving some productive end. For example, in standard principal agent models, high levels of autonomy are warranted when an agent has better information than does the principal about the consequences of actions, and can be rewarded on the basis of the value created by selecting the best action from a choice set. Even in Ellingsen and Johannesson's esteem model, autonomy awarded by the principal serves the instrumental purpose of allowing the agent to signal valuable information to the principal. Economists have only begun to explore the interesting and provocative possibility that autonomy has value in and of itself and that the use of targeted extrinsic rewards (in the psychological sense) undermines an individual's feeling of autonomy and competence.[73]

6. CONCLUSIONS

Our purpose in writing this chapter is to assess the contribution of behavioral economic ideas to the study of agency in employment relationships. In Section 2 we introduce the basic logic of standard agency models and in Section 3 we discuss the complications that arise when incentives must serve "double duty" as is the case where firms have to worry about adverse selection or multi-tasking. In Section 4 we introduce the core behavioral idea of "other regarding preferences" and consider effects on agency relationships of various manifestations of these preferences—equity considerations, effort norms, norms of professional practice and identity. In Section 5 we return to the theme of double duty incentives, and consider the possibility that incentives have the two-fold effect of motivating desired behaviors while also reinforcing (or undermining) intrinsic motives.

The narrow focus of our paper has caused us to give short shrift to many important contributions that behavioral economics has made to our discipline. We say relatively little about such important behavioral economic topics as prospect theory, hyperbolic discounting, mental accounting, status-quo biases and default rules, cognitive dissonance, or bounded rationality. Perhaps more noteworthy than the behavioral issues we have left out of this essay are the standard methodological approaches that we have kept in. Our intention has been to remain theoretically grounded and methodologically conservative. In each section of the paper we represent purposive behavior by analyzing equilibrium behaviors that emerge when individual agents maximize a utility function subject to participation constraints and the constraints imposed by incentive and monitoring systems. Also, consistent with standard economic analysis, we are careful to consider the ways in which equilibrium outcomes are shaped by market competition and by the selection of agents into employment relationships.

Even with this deliberately conservative approach, we find that the introduction of behavioral features into agency models leads to novel and important results: Inequity

[73] See, e.g., Benz and Frey's (2008) research on the value of independence and Dur and Glazer's (2008) work on the desire by workers for impact.

aversion among agents leads to lower powered incentives than would otherwise be the case, but this effect can be undone in certain competitive environments. Effort norms and "gift exchange" can support high effort levels even when monitoring and incentives are problematic, but reliance on effort norms requires that principals be exquisitely attuned to the ways in which their actions influence employee morale. Professional norms can have the effect of protecting consumers from exploitation by professionals and this effect can be reinforced by properly designed incentives. The protective value of these norms can, however, be undermined by self-serving biases that distort the judgement of professionals in unconscious ways. Identity matters for the resolution of agency problems within employment relationships and can help explain important empirical anomalies in labor markets. High powered extrinsic incentives can have a corrosive effect on the motivation of employees, especially when the employees work in "mission driven" or "caring" organizations or when preferences or identity are endogenously shaped by the incentives to which employees are exposed.

The application of behavioral economics to agency in employment relationships is a relatively new area of research. It is worthwhile then to speculate on what might be especially promising areas for future research. We highlight four such areas:

First, given the pivotal importance of professional norms for well functioning markets in health care and financial services, we think it would be useful to investigate more thoroughly the behavioral foundations of conflicts of interest. Very little is known about the ways these conflicts shape the psychology of decision making, and having a clearer understanding of this issue may be quite important for designing efficient and effective regulatory policies.

Second, models of identity have a great deal of appeal, because families, schools, and firms appear to devote enormous resources to shaping and refining the identity of their participants. As currently specified, however, models of identity are so flexible that they may not generate falsifiable conclusions. A satisfactory understanding of the economics of identity will therefore require either a more structured modeling approach or, more likely, the accumulation of additional sociological and psychological data on the nature of identity so that the parameters of the models can be empirically constrained.

Third, much more needs to be learned about the relationship between public policy and income and effort norms. Are Levy and Temin, for example, correct in their assertion that changes in Federal government policy in the 1980's shifted the tolerance for income inequality throughout the labor market? Are Akerlof, Dickens and Perry correct that the effectiveness of monetary policy is determined by the workings of reciprocity and gift exchange in the workplace? At present we do not have definitive answers to these questions.

Finally, although the theory is new and the evidence not yet conclusive, we are intrigued by the notion that extrinsic rewards can undermine intrinsic motives. In health care, corporate governance, education, and other important settings, standard models

typically prescribe some sort of "pay for performance" for resolving agency issues. This prescription must be greatly modified if we can identify people and contexts where high powered financial incentives undermine employee motives to do the right thing.

Clearly, there is much more to discover about the behavioral economics of agency in employment relationships.

REFERENCES

Acemoglu, Daron, Kremer, Michael, Mian, Atif, 2008. Incentives in markets, firms, and governments. Journal of Law, Economics, and Organization 24 (2), 273–306.

Akerlof, George A., 1976. The economics of caste and of the rat race and other woeful tales. Quarterly Journal of Economics 90 (4), 599–617.

Akerlof, George A., 1982. Labor contracts as partial gift exchange. Quarterly Journal of Economics 97 (4), 543–569.

Akerlof, George A., Dickens, William T., Perry, George L., 2000. Near-rational wage and price setting and the long-run Phillips curve. Brookings Papers on Economic Activity 2000 (1), 1–44.

Akerlof, George A., Kranton, Rachel, 2000. Economics and identity. Quarterly Journal of Economics 115 (3), 715–753.

Akerlof, George A., Kranton, Rachel, 2005. Identity and the economics of organizations. Journal of Economic Perspectives 19 (1), 9–32.

Akerlof, George A., Yellen, Janet L., 1985. A near-rational model of the business cycle, with wage and price inertia. Quarterly Journal of Economics 100 (5), 823–838.

Akerlof, George A., Yellen, Janet L., 1990. The fair wage-effort hypothesis and unemployment. Quarterly Journal of Economics 105 (2), 255–283.

Allen, Franklin, Gale, Douglas, 1992. Measurement distortion and missing contingencies in optimal contracts. Economic Theory 2 (1), 1–26.

Allgulin, Magnus, Ellingsen, Tore, 2002. Monitoring and pay. Journal of Labor Economics 20 (2 part 1), 201–216.

Andreoni, James, Douglas Bernheim, B., 2009. Social Image and the 50-50 norm: a theoretical and experimental analysis of audience effect. Econometrica 77 (5), 1607–1636.

Arlen, Jennifer, MacLeod, W. Bentley, 2005. Torts, expertise, and authority: liability of physicians and managed care organizations. RAND Journal of Economics 36 (3), 494–519.

Arrow, Kenneth J., 1998. What has economics to say about racial discrimination? Journal of Economic Perspectives 12 (2), 91–100.

Austen-Smith, David, Fryer Jr., Roland G., 2005. An economic analysis of 'acting white'. Quarterly Journal of Economics 120 (2), 551–583.

Avorn, Jerry, 2004. Powerful Medicines: the Benefits, Risks and Costs of Prescription Drugs. Alfred A. Knopf, New York.

Babcock, Linda, Laschever, Sara, 2003. Women don't ask: negotiation and the gender divide. Princeton University Press, Princeton, Oxford.

Babcock, Linda, Loewenstein, George, 1997. Explaining bargaining impasse: the role of self-serving biases. Journal of Economic Perspectives 11 (1), 109–126.

Babcock, Linda, Loewenstein, George, Issacharoff, Samuel, Camerer, Colin, 1995. Biased judgments of fairness in bargaining. American Economic Review 85 (5), 1337–1343.

Bandiera, Oriana, Barankay, Iwan, Rasul, Imran, 2009. Social connections and incentives in the workplace: evidence from personnel data. Econometrica 77 (July), 1047–1094.

Bebchuk, Lucian, Fried, Jesse M., 2003. Executive compensation as an agency problem. Journal of Economic Perspectives 17 (3), 71–92.

Bebchuk, Lucian, Grinstein, Yaniv, 2005. The growth of executive pay. Oxford Review of Economic Policy 21 (2), 283–303.

Becker, Gary, S., 1957. The Economics of Discrimination. University of Chicago Press, Chicago.

Becker, Gary S., 1981. A Treatise on the Family. NBER Books.

Bem, Daryl J., 1972. Self-perception theory. In: Berkowitz, Leonard (Ed.), Advances in Experimental Social Psychology, Vol. 6, Academic Press, New York.
Bénabou, Roland, Tirole, Jean, 2003. Intrinsic and extrinsic motivation. Review of Economic Studies 70 (3), 489–520.
Bénabou, Roland, Tirole, Jean, 2006. Incentives and prosocial behavior. American Economic Review 96 (5), 1652–1678.
Benz, Matthias, Frey, Bruno S., 2008. Being independent is a great thing: subjective evaluations of self-employment and hierarchy. Economica 75, 362–383.
Berman, Eli, 2000. Sect, subsidy, and sacrifice: an economist's view of ultra-orthodox jews. Quarterly Journal of Economics 115 (3), 905–953.
Bernheim, B. Douglas, 1994. A theory of conformity. Journal of Political Economy 102 (5), 841–877.
Bernheim, B. Douglas, Severinov, Sergei, 2003. Bequests as signals: an explanation for the equal division puzzle. Journal of Political Economy 111 (4), 733–764.
Bertrand, Marianne, Mullainathan, Sendhil, 2001. Are CEOS rewarded for luck? The ones without principals are. Quarterly Journal of Economics 116 (3), 901–932.
Besley, Timothy, Ghatak, Maitreesh, 2005. Competition and incentives with motivated agents. American Economic Review 95 (3), 616–636.
Bewley, Truman F., 1999. Why wages don't fall during a recession. Harvard University Press, Cambridge, MA.
Bewley, Truman F., 2000. Comments and discussion. Brookings Papers on Economic Activity 2000 (1), 45–50.
Bisin, Alberto, Verdier, Thierry, 2000. 'Beyond the melting pot': cultural transmission, marriage, and the evolution of ethnic and religious traits. Quarterly Journal of Economics 115 (3), 955–988.
Black, Dan A., Gates, Gary, Sanders, Seth, Taylor, Lowell J., 2000. Demographics of the gay and lesbian population in the United States: evidence from available systematic evidence. Demography 37 (2), 139–154.
Black, Dan A., Sanders, Seth, Taylor, Lowell J., 2007. The economics of lesbian and gay families. Journal of Economic Perspectives 21 (2), 53–70.
Bowles, Hannah Riley, Babcock, Linda, Lai, Lei, 2006. Social incentives for gender differences in the propensity to initiate negotiations: sometimes it does hurt to ask. Organizational Behavior and Human Decision Processes 103 (1), 84–103.
Bowles, Hannah Riley, Babcock, Linda, McGinn, Kathleen L., 2005. Constraints and triggers: situational mechanics of gender in negotiation. Journal of Personality and Social Psychology 89 (6), 951–965.
Bowles, Samuel, 1985. The production process in a competitive economy: Walrasian, Neo-Hobbesian, and Marxian models. American Economic Review 75 (1), 16–36.
Bowles, Samuel, Gintis, Herbert, 1977. Schooling in Capitalist America: educational Reform and the Contradictions of Economic Life. Basic Books, New York.
Bowles, Samuel, Gintis, Herbert, Osborne, Melissa, 2001. The determinants of earnings: a behavioral approach. Journal of Economic Literature 39 (4), 1137–1176.
Brown, Charles, Medoff, James, 1989. The employer size-wage effect. Journal of Political Economy 97 (5), 1027–1059.
Bulow, Jeremy I., Summers, Lawrence H., 1986. A theory of dual labor markets with application to industrial policy, discrimination, and Keynesian unemployment. Journal of Labor Economics 4 (3), 376–414.
Burdett, Kenneth, Mortensen, Dale T., 1998. Wage differentials, employer size, and unemployment. International Economic Review 39 (2), 257–273.
Cain, Daylian M., Loewenstein, George, Moore, Don A., 2005. The dirt on coming clean: perverse effects of disclosing conflicts of interest. Journal of Legal Studies 34 (1), 1–25.
Camerer, Colin, Babcock, Linda, Loewenstein, George, Thaler, Richard, 1997. Labor supply of New York City cabdrivers: one day at a time. Quarterly Journal of Economics 112 (2), 407–441.
Camerer, Colin, Loewenstein, George, 2004. Behavioral economics: past, present, future. In: Camerer, Colin, Loewenstein, George, Rabin, Matthew (Eds.), Advances in Behavioral Economics. Princeton University Press, Princeton, NJ.
Cappelli, Peter, Chauvin, Keith, 1991. An interplant test of the efficiency wage hypothesis. Quarterly Journal of Economics 106 (3), 769–787.

Cebul, Randall, Rebitzer, James, Taylor, Lowell, Votruba, Mark, 2008. Organizational fragmentation in the US health care system. Journal of Economic Perspectives 22 (4).
Charness, Gary, 2004. Attribution and reciprocity in an experimental labor market. Journal of Labor Economics 22 (3), 665–688.
Charness, Gary, Kuhn, Peter, 2011. Lab labor: what can labor economists learn from the lab? In: Developments in Research Methods and their Application. In: Ashenfelter, Orley, Card, David E. (Eds.), Handbook of Labor Economics, vol. 4a. pp. 229–330. (Chapter 3).
Charness, Gary, Rabin, Matthew, 2002. Understanding social preferences with simple tests. Quarterly Journal of Economics 117 (3), 817–869.
Cho, In-Koo, Kreps, David M., 1987. Signaling games and stable equilibria. Quarterly Journal of Economics 102 (2), 179–221.
Coase, Ronald H., 1937. The nature of the firm. Economica 4 (16), 386–405.
Cooper, David J., Rebitzer, James B., 2006. Managed care and physician incentives: the effects of competition on the cost and quality of care. B.E. Journals in Economic Analysis and Policy: Contributions to Economic Analysis and Policy 5 (1), 1–30.
Cornell, Bradford, Welch, Ivo, 1996. Culture, information, and screening discrimination. Journal of Political Economy 104 (3), 542–571.
Courty, Pascal, Marschke, Gerald, 2004. An empirical investigation of gaming responses to explicit performance incentives. Journal of Labor Economics 22 (1), 23–56.
Dana, Jason, Loewenstein, George, 2003. A social science perspective on gifts to physicians from industry. Journal of the American Medical Association 290 (2), 252–255.
Deci, Edward L., Koestner, Richard, Ryan, Richard M., 1999. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. Psychological Bulletin 125 (6), 627–668.
Deci, Edward L., Ryan, Richard M., 1985. Intrinsic motivation and self-determination in human behavior. Plenum Press, New York.
Delfgaauw, Josse, Dur, Robert, 2007. Signaling and screening of worker motivation. Journal of Economic Behavior and Organization 62, 605–624.
Delfgaauw, Josse, Dur, Robert, 2008. Incentives and workers' motivation in the public sector. Economic Journal 118, 171–191.
DellaVigna, Stefano, 2009. Psychology and economics: evidence from the field. Journal of Economic Literature 47 (2), 315–372.
deQuervain, Dominique J.F., et al., 2004. The neural basis of altruistic punishment. Science 305 (5688), 1254–1258.
Dur, Robert, Glazer, Amihai, 2008. The desire for impact. Journal of Economic Psychology 29 (3), 285–300.
Ekman, Paul, 2001. Telling lies: Clues to Deceit in the Marketplace, Politics, and Marriage, 3rd ed., W. W. Norton and Company, New York City.
Ellingsen, Tore, Johannesson, Magnus, 2007. Paying respect. Journal of Economic Perspective 21 (4), 135–149.
Ellingsen, Tore, Johannesson, Magnus, 2008. Pride and prejudice: the human side of incentive theory. American Economic Review 98 (3), 990–1008.
Encinosa III, William E., Gaynor, Martin, Rebitzer, James B., 2007. The sociology of groups and the economics of incentives: theory and evidence on compensation systems. Journal of Economic Behavior and Organization 62 (2), 187–214.
Falk, Armin, Fehr, Ernst, Zehnder, Christian, 2006. Fairness perceptions and reservation wages—the behavioral effects of minimum wage laws. Quarterly Journal of Economics 121 (4), 1347–1381.
Falk, Armin, Ichino, Andrea, 2006. Clean Evidence on Peer Effects. Journal of Labor Economics 24 (1), 39–57.
Falk, Armin, Kosfeld, Michael, 2006. The hidden costs of control. American Economic Review 96 (5), 1611–1630.
Farber, Henry S., 2005. Is tomorrow another day? The labor supply of New York City cabdrivers. Journal of Political Economy 113 (1), 46–82.

Farber, Henry S., 2008. Reference-dependent preferences and labor supply: the case of New York City taxi drivers. American Economic Review 98 (3), 1069–1082.
Fehr, Ernst, Gaecher, Simon, 2000. Cooperation and punishment in public goods experiments. American Economic Review 90 (4).
Fehr, Ernst, Schmidt, Klaus M., 1999. A theory of fairness, competition, and cooperation. Quarterly Journal of Economics 114 (3), 817–868.
Fehr, Ernst, Goette, Lorenz, 2007. Do workers work more if wages are high? Evidence from a randomized field experiment. American Economic Review 97 (1), 298–317.
Fehr, Ernst, Kirchsteiger, Georg, Riedl, Aron, 1998. Gift exchange and reciprocity in competitive experimental markets. European Economic Review 42 (1), 1–34.
Fernandez, Raquel, Fogli, Alessandra, Olivetti, Claudia, 2004. Mothers and sons: preference formation and female labor force dynamics. Quarterly Journal of Economics 119 (4), 1249–1299.
Festinger, Leon, Carlsmith, James, 1959. Cognitive consequences of force compliance. Journal of Personality and Social Psychology 58 (2), 203–210.
Fischbacher, Uris, Fong, Christina M., Fehr, Ernst, 2009. Fairness, errors and the power of competition. Journal of Economic Behavior and Organization 72 (1), 527–545.
Fischer, Huddart, 2008. Optimal contracting with endogenous social norms. American Economic Review 98 (4), 1459–1475.
Fisman, Raymond, Miguel, Edward, 2007. Corruption, norms, and legal enforcement: evidence from diplomatic parking tickets. Journal of Political Economy 115 (6), 1020–1048.
Frank, Robert H., 1984. Are workers paid their marginal products? American Economic Review 74 (4), 549–571.
Frank, Robert H., 1985. Choosing the Right Pond: Human Behavior and the Quest for Status. Oxford University Press, New York.
Frank, Robert H., 1988. Passions within reason: the strategic role of the emotions. W. W. Norton and Company, New York City.
Frey, Bruno S., 1997. Not Just for The Money: An Economic Theory of Personal Motivation. Edward Elgar, Cheltenham.
Frey, Bruno S., Jegen, Reto, 2001. Motivation crowding theory. Journal of Economic Surveys 15 (5), 589–611.
Frey, Bruno S., Oberholzer-Gee, Felix, 1997. The cost of price incentives: an empirical analysis of motivation crowding-out. American Economic Review 87 (4), 746–755.
Gayle, George-Levi, Miller, Robert A., 2009. Has moral hazard become a more important factor in managerial compensation? American Economic Review 99 (5), 1740–1769.
Gaynor, Martin, Rebitzer, James, Taylor, Lowell J., 2004. Physician incentives in health maintenance organizations. Journal of Political Economy 112 (4).
Genesove, David, Mayer, Christopher, 2001. Loss aversion and seller behavior: evidence from the housing market. Quarterly Journal of Economics 116 (4), 1233–1260.
Gibbons, Robert, 1998. Incentives in organizations. Journal of Economic Perspectives 12 (4), 115–132.
Gibbons, Robert, Waldman, Michael, 1999. A theory of wage and promotion dynamics inside firms. Quarterly Journal of Economics 114 (4), 1321–1358.
Gicheva, Dora, 2009. Working Long Hours and Career Wage Growth. Working Paper, Yale University.
Gintis, Herbert, Bowles, Samuel, Boyd, Robert, Fehr, Ernst, 2003. Explaining altruistic behavior in humans. Evolution and Human Behavior 24 (3), 153–172.
Gneezy, Uri, Muriel, Niederle, Aldo, Rustichini, 2003. Performance in competitive environments: gender differences. Quarterly Journal of Economics 118 (3), 1049–1074.
Gneezy, Uri, Rustichini, Aldo, 2000a. A fine is a price. Journal of Legal Studies 29 (1), 1–17.
Gneezy, Uri, Rustichini, Aldo, 2000b. Pay enough or don't pay at all. Quarterly Journal of Economics 115 (3), 791–810.
Grogger, Jeffrey, 2009. Welfare reform, returns to experience, and wages: using reservation wages to account for sample selection biases. Review of Economics and Statistics 91 (3), 490–502.
Grossman, Sanford J., Hart, Oliver D., 1986. The costs and benefits of ownership: a theory of vertical and lateral integration. Journal of Political Economy 94 (4), 691–719.

Groves, Theodore, Hong, Yongmiao, McMillan, John, Naughton, Barry, 1994. Autonomy and incentives in Chinese state enterprises. Quarterly Journal of Economics 109 (1), 183–209.
Groves, Theodore, Hong, Yongmiao, McMillan, John, Naughton, Barry, 1995. China's evolving managerial labor market. Journal of Political Economy 103 (4), 873–892.
Hall, Brian, Knox, Thomas, 2004. Underwater options and the dynamics of executive pay-to-performance sensitivities. Journal of Accounting Research 42 (2), 365–412.
Hall, Brian, Liebman, Jeffrey, 1998. Are CEOS really paid like bureaucrats? Quarterly Journal of Economics 113 (3), 653–691.
Hall, Brian J., Murphy, Kevin J., 2003. The trouble with stock options. The Journal of Economic Perspectives 17 (3), 49–70.
Heron, Randall, Lie, Erik, 2009. What fraction of stock option grants to top executives have been backdated or manipulated? Management Science 55 (4), 513–525.
Heyes, Anthony, 2005. The economics of vocation or why is a badly paid nurse a good nurse? Journal of Health Economics 24, 561–569.
Holmström, Bengt, 1979. Moral hazard and observability. Bell Journal of Economics 10 (1), 74–91.
Holmström, Bengt, 1999. Managerial incentive problems: a dynamic perspective. Review of Economic Studies 66 (1), 169–182.
Holmström, Bengt, Milgrom, Paul, 1991. Multitask principal-agent analyses: incentive contracts, asset ownership, and job design. Journal of Law, Economics, and Organization 7, 24–52. Special Issue (Papers from the Conference on the New Science of Organization).
Holmström, Bengt, Milgrom, Paul, 1994. The firm as an incentive system. American Economic Review 84 (4), 972–991.
Hornstein, Andreas, Krusell, Per, Violante, Giovanni, 2007. Frictional wage dispersion in search models: a quantitative assessment. NBER Working Paper.
Ichino, Andrea, Maggi, Giovanni, 2000. Work Environment and individual background: explaining regional shirking differentials in a large Italian firm. Quarterly Journal of Economics 115 (3), 1057–1090.
Jackson, Howell E., 2008. The trilateral dilemma in financial regulation. In: Maria Lusardi, Anna (Ed.), Improving the Effectiveness of Financial Education and Savings Programs. University of Chicago Press, Chicago.
Jackson, Kirabo C., Bruegmann, Elias, 2009. Teaching students and teaching each other: the importance of peer learning for teachers. Manuscript.
Jacob, Brian A., 2005. Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago public schools. Journal of Public Economics Volume 89 (5–6), 761–796.
Jacob, Brian A., Levitt, Steven D., 2003. Rotten apples: an investigation of the prevalence and predictors of teacher cheating. Quarterly Journal of Economics 118 (3), 843–877.
Jensen, Michael, Murphy, Kevin J., 1990. Performance pay and top-management incentives. Journal of Political Economy 98, 225–264.
Kahn, Lawrence M., Sherer, Peter D., 1990. Contingent pay and managerial performance. Industrial and Labor Relations Review 43 (3), 107S-120S.
Kandel, Eugene, Lazear, Edward P., 1992. Peer pressure and partnerships. Journal of Political Economy 100 (4), 801–817.
Knez, Marc, Simester, Duncan, 2001. Firm-wide incentives and mutual monitoring at continental airlines. Journal of Labor Economics 19 (4), 743–772.
Krueger, Alan B., Mas, Alexandre, 2004. Strikes, scabs, and tread separations: labor strife and the production of defective Bridgestone/Firestone tires. Journal of Political Economy 112 (2), 253–289.
Krueger, Alan B., Summers, Lawrence H., 1988. Efficiency wages and the inter-industry wage structure. Econometrica 56 (2), 259–293.
Landers, Renee, Rebitzer, James, Taylor, Lowell J., 1996. Rat race redux: adverse selection in the determination of work hours in law firms. American Economic Review 86 (3).
Landers, Renee, Rebitzer, James, Taylor, Lowell J., 1997. Work norms in professional labor markets. In: Blau, Francine, Ehrenberg, Ronald (Eds.), Gender and Family Issues in the Workplace. Russell Sage Press, New York.
Lang, Kevin, 1986. A language theory of discrimination. Quarterly Journal of Economics 101, 363–382.
Lazear, Edward P., 1989. Pay equality and industrial politics. Journal of Political Economy 97 (3), 561–580.

Lazear, Edward P., 1998. Personnel economics for managers. John Wiley and Sons, New York.
Lazear, Edward P., 1999. Culture and language. Journal of Political Economics 107 (6, part 2), S95–S126.
Lazear, Edward P., 2000. Performance pay and productivity. American Economic Review 90 (5), 1346–1361.
Lepper, Mark R., Greene, David, Nisbett, Richard, 1973. Undermining children's intrinsic interest with extrinsic reward: a test of the 'overjustification' hypothesis. Journal of Personality and Social Psychology 28 (1), 129–137.
Levine, David I., 1991. Just-cause employment policies in the presence of worker adverse selection. Journal of Labor Economics 9 (3), 294–305.
Levy, Frank, Temin, Peter, 2007. Inequality and Institutions in 20th Century America. NBER Working Paper.
MacLeod, W. Bentley, Malcomson, James, 1989. Implicit contracts, incentive compatibility, and involuntary unemployment. Econometrica 57, 447–480.
Malcomson, James M., 1984. Work incentives, hierarchy, and internal labor markets. Journal of Political Economy 92 (3), 486–507.
Malcomson, James M., 1999. Individual employment contracts. In: Ashenfelter, Orley, Card, David (Eds.), Handbook of Labor Economics, vol. 3 (number 3).
Mas, Alexandre, 2006. Pay, reference points, and police performance. Quarterly Journal of Economics 121 (3), 783–821.
Mas, Alexandre, 2008. Labour unrest and the quality of production: evidence from the construction equipment resale market. Review of Economic Studies 75 (1), 229–258.
Mas, Alexandre, Moretti, Enrico, 2009. Peers at work. American Economic Review 99 (1), 112–145.
McCabe, Kevin A., Rigdon, Mary L., Smith, Vernon L., 2003. Positive reciprocity and intentions in trust games. Journal of Economic Behavior and Organization 52 (2), 267–275.
McMillan, John, Whalley, John, Zhu, Lijing, 1989. The impact of China's economic reforms on agricultural productivity growth. Journal of Political Economy 97 (4), 781–807.
Moore, Don A., Lowenstein, George, 2004. Self-interest, automaticity, and the psychology of conflict of interest. Social Justice Research 17 (2), 189–202.
Nagin, Daniel, Rebitzer, James, Sanders, Seth, Taylor, Lowell, 2002. Monitoring and motivation: an analysis of a field experiment. American Economic Review 92 (4).
Neal, Derek, 2006. Why has black-white convergence stopped? In: Hanushek, Eric, Welch., Finis (Eds.), Handbook of the Economics of Education, vol. 1. Elsevier, Amsterdam.
Niederle, Muriel, Vesterlund, Lise, 2007. Do women shy away from competition? Do men compete too much? Quarterly Journal of Economics 122 (3), 1067–1101.
Oyer, Paul, 1998. The Fiscal year ends and nonlinear incentive contracts: the effect on business seasonality. Quarterly Journal of Economics 113 (1), 14985.
Oyer, Paul, Shaefer, Scott, 2011. Personnel economics: hiring and incentives. In: New Developments and Research on Labor Markets. In: Ashenfelter, Orley, Card, David E. (Eds.), Handbook of Labor Economics, vol. 4b. pp. 1769–1823. (Chapter 20).
Prendergast, Canice, 1999. The provision of incentives in firms. Journal of Economic Literature 37 (1), 7–63.
Prendergast, Canice, 2007. The motivation and bias of bureaucrats. American Economic Review 97 (1), 180–196.
Prendergast, Canice, Topel, Robert H., 1996. Favoritism in organizations. Journal of Political Economy 104 (5), 958–978.
Rabin, Matthew, 1993. Incorporating fairness into game theory and economics. American Economic Review 83, 1281–1302.
Rabin, Matthew, Thaler, Richard H., 2001. Anomalies: risk aversion. Journal of Economic Perspectives 15 (1), 13.
Rebitzer, James B., 1995. Job safety and contract workers in the petrochemical industry. Industrial Relations 34 (1), 40–57.
Rebitzer, James B., Taylor, Lowell J., 1991. A model of dual labor markets when product demand is uncertain. Quarterly Journal of Economics 106 (4), 1373–1383.
Rebitzer, James B., Taylor, Lowell J., 1995a. Efficiency wages and employment rents: the employer size wage effect in the job market for lawyers. Journal of Labor Economics 13 (4), 678–708.

Rebitzer, James B., Taylor, Lowell J., 1995b. The consequences of minimum wage laws: some new theoretical ideas. Journal of Public Economics 56 (2), 245–255.
Ritter, Joseph A., Taylor, Lowell J., 1994. Workers as creditors: efficiency wages and performance bonds. American Economic Review 84 (3), 694–704.
Ritter, Joseph A., Taylor, Lowell J., Racial Disparity in Unemployment. Review of Economics and Statistics (forthcoming).
Rousseau, Denise M., 1995. Psychological Contracts in Organizations: Understanding Written and Unwritten Agreements. Sage Publications, Thousand Oaks.
Rudman, P., Glick, L.A., 1999. Feminized management and backlash toward agentic women: the hidden costs to women of a kinder, gentler image of middle managers. Journal of Personality and Social Psychology 77 (5), 1004–1010.
Ryan, Richard M., Deci, Edward L., 2000. Intrinsic and extrinsic motivations: classic definitions and new directions. Contemporary Education Psychology 25 (1), 54–67.
Sally, David, 2002. Two economic applications of sympathy. Journal of Law, Economics, and Organization 18 (2), 455–487.
Shapiro, Carl, Stiglitz, Joseph, 1984. Involuntary unemployment as a worker discipline device. American Economic Review 74 (3), 433–444.
Simon, Herbert A., 1951. A formal theory of the employment relationship. Econometrica 19 (3), 293–305.
Sliwka, Dirk, 2007. Trust as a signal of a social norm and the hidden costs of incentive schemes. American Economic Review 97 (3), 999–1012.
Small, Deborah, Gelfand, Michele, Babcock, Linda, Gettman, Hilary, 2007. Who goes to the bargaining table? The influence of gender and framing on the initiation of negotiation. Journal of Personality and Social Psychology 93 (4), 600–613.
Spier, Kathryn E., 1992. Incomplete contracts and signalling. RAND Journal of Economics 23 (3), 432–443.
Stevenson, Betsy, Wolfers, Justin, 2008. Economic growth and subjective well-being: reassessing the Easterlin paradox. Brookings Papers on Economic Activity.
Taylor, Lowell J., 2007. Optimal wages in the market for nurses: an analysis based on Heyes' model. Journal of Health Economics 26 (5), 1027–1030.
Thaler, Richard H., Sunstein, Cass R., 2008. Nudge: Improving Decisions About Health, Wealth, and Happiness. Yale University Press.
Williamson, Oliver E., 1985. The Economic Institutions of Capitalism. Free Press, New York.
Valley, Kathleen L., Moag, Joseph, Bazerman, Max H., 1998. 'A matter of trust:' Effects of communication on the efficiency and distribution of outcomes. Journal of Economic Behavior and Organization 34 (2), 211–238.