The Journal of Political Economy Vol. LXX, No.5, Part 2, 1962

INVESTMENT IN HUMAN CAPITAL: A THEORETICAL ANALYSIS[1]

GARY S. BECKER
Columbia University and National Bureau of Economic Research

I. INTRODUCTION

Some activities primarily affect future well-being, while others have their main impact in the present. Dining is an example of the latter, while purchase of a car exemplifies the former. Both earnings and consumption can be affected: on-the-job training primarily affects earnings, a new sail boat primarily affects consumption, and a college education is said to affect both. The effects may operate either through physical resources, such as a sail boat, or through human resources, such as a college education. This paper is concerned with activities that influence future real income through the imbedding of resources in people. This is called investing in human capital.

The many ways to invest include schooling, on-the-job training, medical care, vitamin consumption, and acquiring information about the economic system. They differ in the relative effects on earnings and consumption, in the amount of resources typically invested, in the size of returns, and in the extent to which the connection between investment and return is perceived. But all improve the physical and mental abilities of people and thereby raise real income prospects.

People differ substantially in their economic well-being, both among countries and among families within a given country. For a while economists were relating these differences primarily to differences in the amount of physical capital since richer people had more physical capital than others. It has become increasingly evident, however, from studies of income growth[2] that factors other than physical resources play a larger role than formerly believed, thus focusing attention on less tangible resources, like the knowledge possessed. A concern with investment in human capital, therefore, ties in closely with the new emphasis on intangible resources and may be useful in attempts to understand the inequality in income among people.

The original aim of my study was to estimate the money rate of return to college and high-school education in the United States. In order to set these estimates in proper context I undertook a brief formulation of the theory of investment in human capital. It soon became clear to me, however, that more than a restatement was called for: while important and pioneering work had been done on the economic return to various

[1] I am greatly indebted to the Carnegie Corporation of New York for the support given to the National Bureau of Economic Research to study investment in education and other kinds of human capital. I benefited greatly from many discussions with my colleague Jacob Mincer, and also with other participants in the Labor Workshop of Columbia University. Although many persons offered valuable comments on the draft prepared for the conference, I am especially indebted to the detailed comments of Theodore Schultz, George Stigler, and Shirley Johnson.

[2] The evidence for the United States appears to show that the growth in capital per capita explains only a small part of the growth in per capita income and that the growth in "technology" explains most of it. On this see S. Fabricant, *Economic Progress and Economic Change: 34th Annual Report of the National Bureau of Economic Research* (New York: National Bureau of Economic Research, 1954).

occupations and education classes,[3] there have been few, if any, attempts to treat the process of investing in people from a general viewpoint or to work out a broad set of empirical implications. I began then to prepare a general analysis of investment in human capital.

As the work progressed, it became clearer and clearer that much more than a gap in formal economic analysis would be filled, for the analysis of human investment offered a unified explanation of a wide range of empirical phenomena which had either been given *ad hoc* interpretations or had baffled investigators. Among these are the following: (1) Earnings typically increase with age at a decreasing rate. Both the rate of increase and the rate of retardation tend to be positively related to the level of skill. (2) Unemployment rates tend to be negatively related to the level of skill. (3) Firms in underdeveloped countries appear to be more "paternalistic" toward employees than those in developed countries. (4) Younger persons change jobs more frequently and receive more schooling and on-the-job training than older persons do. (5) The distribution of earnings is positively skewed, especially among professional and other skilled workers. (6) Abler persons receive more education and other kinds of training than others. (7) The division of labor is limited by the extent of the market. (8) The typical investor in human capital is more impetuous and thus more likely to err than is the typical investor in tangible capital. What a diverse and possibly even confusing array! Yet all these as well as many other important empirical implications can be derived from very simple theoretical arguments. The purpose of this paper is to set out these arguments in some generality, with the emphasis placed on empirical implications, although little empirical material is presented. My own empirical work will appear in a later study.

First, a lengthy discussion of on-the-job training is presented and then, much more briefly, discussions of investment in schooling, information, and health. On-the-job training is dealt with so elaborately not because it is more important than other kinds of investment in human capital—although its importance is often underrated—but because it clearly illustrates the effect of human capital on earnings, employment, and other economic variables. For example, the close connection between foregone and direct costs or the effect of human capital on earnings at different ages is vividly brought out. The extended discussion of on-the-job training paves the way for much briefer discussions of other kinds of investment in human beings.

II. DIFFERENT KINDS OF INVESTMENT

A. ON THE JOB

Theories of firm behavior, no matter how they differ in other respects, almost invariably ignore the effect of the productive process itself on worker productivity. This is not to say that no one recognizes that productivity is affected by the job itself; but the recognition has not been formalized, incorporated into economic analysis, and its implications worked out. We now intend to do just

[3] In addition to the earlier works of Smith, Mill, and Marshall, see H. Clark, *Life Earnings in Selected Occupations in the U.S.* (New York: Harper & Bros., 1937); J. R. Walsh, "Capital Concept Applied to Man," *Quarterly Journal of Economics*, February, 1935; M. Friedman and S. Kuznets, *Income from Independent Professional Practice* (New York: National Bureau of Economic Research, 1945); G. Stigler and D. Blank, *The Demand and Supply of Scientific Personnel* (New York: National Bureau of Economic Research, 1957); and T. W. Schultz, "Investment in Man: An Economist's View," *Social Service Review*, June, 1959.

that, placing special emphasis on the broader economic implications.

Many workers increase their productivity by learning new skills and perfecting old ones while on the job. For example, the apprentice usually learns a completely new skill while the intern develops skills acquired in medical school, and both are more productive afterward. On-the-job training, therefore, is a process that raises future productivity and differs from school training in that an investment is made on the job rather than in an institution that specializes in teaching. Presumably, future productivity can be improved only at a cost, for otherwise there would be an unlimited demand for training. Included in cost are a value placed on the time and effort of trainees, the "teaching" provided by others, and the equipment and materials used. These are costs in the sense that they could have been used in producing current output if they were not used in raising future output. The amount spent and the duration of the training period depend partly on the type of training—more is spent for a longer time on an intern than on an operative—partly on production possibilities, and partly on the demand for different skills.

Each employee is assumed to be hired for a specified time period (in the limiting case this period approaches zero), and for the moment both labor and product markets are assumed to be perfectly competitive. If there were no on-the-job training, wage rates would be given to the firm and would be independent of its actions. A profit-maximizing firm would be in equilibrium when marginal products equaled wages, that is, when marginal receipts equaled marginal expenditures. In symbols

$$MP = W, \qquad (1)$$

where W equals wages or expenditures and MP equals the marginal product or receipts. Firms would not worry too much about the relation between labor conditions in the present and future partly because workers were only hired for one period, and partly because wages and marginal products in future periods would be independent of a firm's current behavior. It can therefore legitimately be assumed that workers have unique marginal products (for given amounts of other inputs) and wages in each period, which are, respectively, the maximum productivity in all possible uses and the market wage rate. A more complete set of equilibrium conditions would be the set

$$MP_t = W_t, \qquad (2)$$

where t refers to the tth period. The equilibrium position for each period would depend only on the flows during that period.

These conditions are altered when account is taken of on-the-job training and the connection thereby created between present and future receipts and expenditures. Training might lower current receipts and raise current expenditures, yet firms could profitably provide this training if future receipts were sufficiently raised or future expenditures sufficiently lowered. Expenditures during each period need not equal wages, receipts need not equal the maximum possible productivity, and expenditures and receipts during all periods would be interrelated. The set of equilibrium conditions summarized in equation (2) would be replaced by an equality between the *present values* of receipts and expenditures. If E_t and R_t represent expenditures and receipts during period t, and i the market discount rate, then the equilibrium condition can be written as

$$\sum_{t=0}^{n-1} \frac{R_t}{(1+i)^{t+1}} = \sum_{t=0}^{n-1} \frac{E_t}{(1+i)^{t+1}}, \quad (3)$$

where n represents the number of periods, and R_t and E_t depend on all other receipts and expenditures. The equilibrium condition of equation (2) has been generalized, for if marginal product equals wages in each period, the present value of the marginal product stream would have to equal the present value of the wage stream. Obviously, however, the converse need not hold.

If training were given only during the initial period, expenditures during the initial period would equal wages plus the outlay on training, expenditures during other periods would equal wages alone, and receipts during all periods would equal marginal products. Equation (3) becomes

$$MP_0 + \sum_{t=1}^{n-1} \frac{MP_t}{(1+i)^t} = W_0 + k + \sum_{t=1}^{n-1} \frac{W_t}{(1+i)^t}, \quad (4)$$

where k measures the outlay on training.

If a new term is defined,

$$G = \sum_{t=1}^{n-1} \frac{MP_t - W_t}{(1+i)^t}, \quad (5)$$

equation (4) can be written as

$$MP_0 + G = W_0 + k. \quad (6)$$

Since the term k only measures the actual outlay on training it does not entirely measure training costs, for excluded is the time that a person spends on this training, time that could have been used to produce current output. The difference between what could have been produced, call this MP_0' and what is produced, MP_0, is the opportunity cost of the time spent in training. If C is defined as the sum of opportunity costs and outlays on training, (6) becomes

$$MP_0' + G = W_0 + C. \quad (7)$$

The term G, the excess of future receipts over future outlays, is a measure of the return to the firm from providing training; and, therefore, the difference between G and C measures the difference between the return from, and the cost of, training. Equation (7) shows that marginal product would equal wages in the initial period only when the return equals costs, or $G = C$; it would be greater or less than wages as the return was smaller or greater than costs. Those familiar with capital theory might argue that this generalization of the simple equality between marginal product and wages is spurious because a full equilibrium would require equality between the return from an investment—in this case, made on the job—and costs. If this implied that $G = C$, marginal product would equal wages in the initial period. There is much to be said for the relevance of a condition equating the return from an investment with costs, but such a condition does not imply that $G = C$ or that marginal product equals wages. The following discussion demonstrates that great care is required in the application of this condition to on-the-job investment.

1. *General.*—Our treatment of on-the-job training produced some general results—summarized in equations (3) and (7)—of wide applicability, but more concrete results require more specific assumptions. In this and the following section two types of on-the-job training are discussed in turn: general and specific. General training is useful in many firms in addition to the firm providing it, as a machinist trained in the army finds his skills of value in steel and aircraft firms,

or a doctor trained (interned) at one hospital finds his skills useful at other hospitals. Most on-the-job training presumably increases the future marginal product of workers in the firm providing it, but general training would also increase their marginal product in many other firms as well. Since in a competitive labor market the wage rates paid by any firm are determined by marginal productivities in other firms, future wage rates as well as marginal products would increase to firms providing general training. These firms could capture some of the return from training only if their marginal product rose by more than their wages. "Perfectly general" training would be equally useful in many firms and marginal products would rise by the same extent in all of them. Consequently, wage rates would rise by exactly the same amount as the marginal product and the firms providing such training could not capture any of the return.

Why, then, do rational firms in competitive labor markets provide general training, for why provide training that brings no return? The answer is that firms would provide general training only if they did not have to pay any of the costs. Persons receiving general training would be willing to pay these costs since training raises their future wages. Hence the cost as well as the return from general training would be borne by trainees, not by firms.

These and other implications of general training can be more formally demonstrated with equation (7). Since wages and marginal products are raised by the same amount, MP_t must equal W_t for all $t = 1, \ldots n-1$, and therefore

$$G = \sum_{t=1}^{n-1} \frac{MP_t - W_t}{(1+i)^t} = 0 . \qquad (8)$$

Equation (7) is reduced to

$$MP_0' = W_0 + C , \qquad (9)$$

or

$$W_0 = MP_0' - C . \qquad (10)$$

In terms of actual marginal product

$$MP_0 = W_0 + k , \qquad (9')$$

or

$$W_0 = MP_0 - k . \qquad (10')$$

The wage of trainees would not equal their opportunity marginal product but would be less by the total cost of training. In other words, employees would pay for general training by receiving wages below their current (opportunity) productivity. Equation (10) has many other implications, and the rest of this section is devoted to developing the more important ones.

Some might argue that a really "net" definition of marginal product obtained by subtracting training costs from "gross" marginal product must equal wages even for trainees. Such an interpretation of net productivity could formally save the equality between marginal product and wages here, but later I show (pp. 18–25) that it cannot always be saved. Moreover, regardless of which interpretation is used, training costs would have to be included in any study of the relation between wages and productivity.

Employees pay for general on-the-job training by receiving wages below what could be received elsewhere. "Earnings" during the training period would be the difference between an income or flow term, potential marginal product, and a capital or stock term, training costs, so that the capital and income accounts would be closely intermixed, with changes in either affecting wages. In other words, earnings of persons receiving on-the-job training would be net of

investment costs and would correspond to the definition of *net* earnings used throughout this paper, which subtracts all investment costs from "gross" earnings. Therefore, our departure with this definition of earnings from the accounting conventions used for transactions in material goods—which separate income from capital accounts to prevent a transaction in capital from *ipso facto*[4] affecting the income side—is not capricious but is grounded in a fundamental difference between the way investment in material and human capital are "written off." The underlying cause of this difference undoubtedly is the widespread reluctance to treat people as capital and the accompanying tendency to treat all wage receipts as earnings.

Intermixing the capital and income accounts could make the reported "incomes" of trainees unusually low and perhaps negative, even though their long-run or lifetime incomes were well above average. Since a considerable fraction of young persons receive some training, and since trainees would tend to have lower current and higher subsequent earnings than other youth, the correlation between current consumption and current earnings of young people[5] would not only be much weaker than the correlation with long-run earnings, but the signs of these correlations might even differ.[6]

Doubt has been cast on the frequent assertion that no allowance is made in the income accounts for depreciation on human capital.[7] A depreciation-type item is deducted, at least from the earnings due to on-the-job training, for the cost would be deducted during the training period. Depreciation on tangible capital does not bulk so large in any one period because it is usually "written off" or depreciated during a period of time designed to approximate its economic life. Hence human and tangible capital appear to differ more in the time pattern of depreciation than in its existence,[8] and the effect on wage income of a rapid "write-off" of human capital is what should often be emphasized and studied.

Our point can be put differently and more rigorously. The ideal depreciation on a capital asset during any period would equal its change in value during the period. In particular, if value rose, a negative depreciation term would have

[4] Of course, a shift between assets having different productivities would affect the income account on material goods even with current accounting practices.

[5] I say "young people" rather than "young families" because as J. Mincer has shown (in a paper to be published in a National Bureau of Economic Research conference volume on labor economics), the labor-force participation of wives is positively correlated with the difference between husbands' long-run and current income. Participation of wives, therefore, makes the correlation between a family's current and a husband's long-run income greater than that between a husband's current and long-run income.

[6] A difference in signs is impossible in Friedman's analysis of consumer behavior because he assumes that transitory and long-run (that is, permanent) incomes are uncorrelated (see his *A Theory of the Consumption Function* [Princeton, N.J.: Princeton University Press, 1959]); we are suggesting that they may be *negatively* correlated for young persons.

[7] See, for example, A. Marshall, *Principles of Economics* (8th ed.; New York: Macmillan Co., 1949); C. Christ, "Patinkin on Money, Interest, and Prices," *Journal of Political Economy*, August, 1957, p. 352; and W. Hamburger, "The Relation of Consumption to Wealth and the Wage Rate," *Econometrica*, January, 1955.

[8] In a recent paper, R. Goode has argued (see "Educational Expenditures and the Income Tax," in Selma J. Mushkin [ed.], *Economics of Higher Education* [Washington: United States Department of Health, Education, and Welfare (forthcoming)]) that educated persons should be permitted to subtract from income a depreciation allowance on tuition payments. Such an allowance is apparently not required for on-the-job training costs; indeed, one might argue, on the contrary, that too much or too rapid depreciation is permitted on such investment.

to be subtracted or a positive appreciation term added to the income from the asset. Since training costs would be deducted from earnings during the training period, the economic "value" of a trainee would at first increase rather than decrease with age, and only later would it begin to decrease.[9]

Training has an important effect on the relation between earnings and age. Suppose that untrained persons received the same earnings regardless of age, as shown by the horizontal line UU in Figure 1. Trained persons would receive lower earnings during the training period because training is paid for then, and higher earnings at later ages because the return is collected then. The combined effect of paying for and collecting the return from training in this way would be to make the age earnings curve of trained persons, shown by TT in Figure 1, steeper than that of untrained persons, the difference being greater the greater the cost of, and return from, the investment.

Not only does training make the curve steeper but, as indicated by Figure 1, also more concave; that is, the rate of increase in earnings is affected more at younger than at older ages. Suppose, to take an extreme case, that training raised the level of marginal productivity but had no effect on the slope, so that the marginal productivity of trained persons was also independent of age. If earnings equaled marginal product, TT would merely be parallel to and higher than UU, showing neither slope nor concavity. Since, however, earnings of trained persons would be below marginal productivity during the training period and equal afterwards, they would rise sharply at the end of the training period and then level off (as shown by the dashed line $T'T'$ in Fig. 1), imparting a concave appearance to the curve as a whole. In this extreme case an extreme concavity appears; in less extreme cases the principle would be the same and the concavity more continuous.

Foregone earnings are an important, although neglected, cost of much human capital and should be treated on the same footing as direct outlays. Indeed, *all* costs appear as foregone earnings to workers

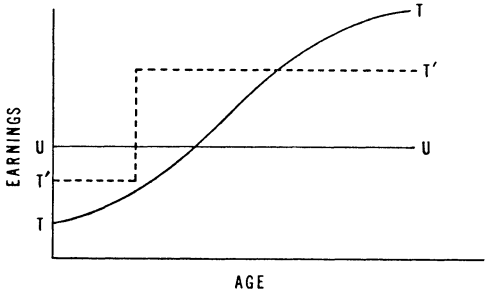

FIG. 1

receiving on-the-job training; that is, all costs appear as lower earnings than could be received elsewhere, although direct outlays, C, may really be an important part of costs. The arbitrariness of the division between foregone and direct costs and the resulting advantage of treating total costs as a whole[10] can be

[9] In my study for the National Bureau of Economic Research I try to measure the relation between depreciation and age for several education classes.

[10] The equivalence between foregone and direct costs applies to consumption as well as to investment decisions. A household can be assumed to maximize a utility function

$$U(X_1, X_2, \ldots X_r),$$

$X_1, \ldots X_r$ being consumption goods, subject to the constraint

$$\sum_{i=1}^{r} p_i X_i = W\left(h - \sum_{j=1}^{r} h_j X_j\right) + y,$$

where p_i is the market price of the ith good, W the average wage rate, y non-wage income, h the total

further demonstrated by contrasting school and on-the-job training. Usually only the direct cost of school training is emphasized, even though the foregone cost is sometimes (as with college education) an important part of the total. A shift of training from schools to on the job would, however, reverse the emphasis and make all costs appear as foregone earnings, even when direct outlays were important.

Income maximizing firms in competitive labor markets would not pay the cost of general training and would pay trained persons the market wage. If, however, training costs were paid, many persons would seek training, few would quit during the training period, and labor costs would be relatively high. Firms that did not pay trained persons the market wage would have difficulty satisfying their skill requirements and would also tend to be less profitable than other firms. Firms that both paid for training and less than the market wage for trained persons would have the worst of both worlds, for they would attract too many trainees and too few trained persons.

These principles have been clearly demonstrated during the last few years in discussions of problems in recruiting military personnel. The military offers training in a wide variety of skills and many—such as piloting and machine repair—are very useful in the civilian sector. Training is provided during part or all of the first enlistment period and used during the remainder of the first period and hopefully during subsequent periods. This hope, however, is thwarted by the fact that re-enlistment rates tend to be inversely related to the amount of civilian-type skills provided by the military.[11] Persons with these skills leave the military more readily because they can receive much higher wages in the civilian sector. Net military wages for those receiving training are higher relative to civilian wages during the first than during subsequent enlistment periods because training costs are largely paid by the military. Not surprisingly, therefore, first-term enlistments for skilled jobs are obtained much more easily than are re-enlistments.

The military is a conspicuous example of an organization that both pays at least part of training costs and does not pay market wages to skilled personnel. It has had, in consequence, relatively easy access to "students" and heavy losses of "graduates." Indeed, its graduates make up the predominate part of the supply in several civilian occupations. For example, well over 90 per cent of United States commercial airline pilots received much of their training in the armed forces. The military, of course, is not a commercial organization judged by profits and losses and has had no difficulty surviving and even thriving.

What about the old argument that

---

number of hours available for either consumption or work, and h_j the number of hours required to consume a unit of the jth good. By transposing terms the constraint can be written as

$$\Sigma(p_i + Wh_i)X_i = Wh + y .$$

The total cost or price of consuming a unit of the ith good is the sum of two components: the market price or direct outlay per unit, p_i, and the foregone earnings per unit, Wh_i. I expect to show in another paper that this formulation of household decisions gives extremely useful insights into a number of important economic problems, such as the choice between labor and "leisure," the effect of price control on prices, the role of queues, and the cause of differences among income classes in price elasticities of demand.

[11] See *Manpower Management and Compensation* (Washington: Government Printing Office, 1957), Vol. I, Chart 3, and the accompanying discussion. The military not only wants to eliminate the inverse relation but apparently would like to create a strong positive relation because they have such a large investment in heavily trained personnel (see *ibid.*).

firms in competitive labor markets have no incentive to provide on-the-job training because trained workers would be bid away by other firms? Firms that train workers are supposed to impart external economies to other firms because the latter can use these workers free of any training charge. An analogy with research and development is often drawn since a firm developing a process that cannot be patented or kept secret would impart external economies to competitors.[12] This argument and analogy would apply if firms were to pay training costs, for they would suffer a "capital loss" whenever trained workers were bid away by other firms. Firms can, however, shift training costs to trainees and have an incentive to do so when faced with competition for their services.

The difference between investment in training and in research and development can be put very simply. Without patents or secrecy, firms in competitive industries cannot establish property rights in innovations, and these innovations become fair game for all comers. Patent systems try to establish these rights so that incentives can be provided to invest in research. Property rights in skills, on the other hand, are automatically vested, for a skill cannot be used without permission of the person possessing it. This property right in skills is the source of the incentive to invest in training and explains why an analogy with unowned innovations is misleading.

2. *Specific.*—Completely general training increases the marginal productivity of trainees by exactly the same amount in firms providing the training as in other firms. Clearly some kinds of training increase productivity by a different amount in firms providing the training than in other firms. Training that increases productivity more in firms providing it will be called specific training. Completely specific training can be defined as training that has no effect on the productivity of trainees that would be useful in in other firms. Much on-the-job training is neither completely specific not completely general but increases productivity more in firms providing it and falls within the definition of specific training. The rest increases productivity by at least as much in other firms and falls within a definition of general training. The previous section discussed general training and this one will cover specific training. A few illustrations of the scope of specific training are presented before a formal analysis is developed.

The military offers some forms of training that are extremely useful in the civilian sector, as already noted. Training is also offered that is only of minor use to civilians: astronauts, fighter pilots, and missile men all illustrate this to a greater or lesser extent. Such training falls within the scope of specific training because productivity is raised in the military but not (much) elsewhere.

Resources are usually spent by firms in familiarizing new employees with their organization,[13] and the knowledge so acquired is a form of specific training because productivity is raised more in the firms acquiring the knowledge than in other firms. Other kinds of hiring costs, such as employment agency fees, the expenses incurred by new employees in finding jobs (what Stigler calls in his paper in this Supplement the "costs of

[12] These arguments can be found in Marshall, *op. cit.*, pp. 565–66, although he compares training to land-tenure systems.

[13] To judge by a sample of firms recently analyzed, formal orientation courses are quite common, at least in large firms (see H. F. Clark and H. S. Sloan, *Classrooms in the Factories* [New York: New York University Press, 1955], chap. iv).

search"), or the time employed in interviewing, testing, checking references, and in bookkeeping do not so obviously raise the knowledge of new employees, but they too are a form of specific investment in human capital, although not training. They are an investment because outlays over a short period create distributed effects on productivity; they are specific because productivity is raised primarily in the firms making the outlays; they are in human capital because they lose their value whenever employees leave. In the rest of this section I usually refer only to on-the-job specific training even though the analysis applies to all on-the-job specific investment.

Even after hiring costs are incurred, firms usually know only a limited amount about the ability and potential of new employees. They try to increase their knowledge in various ways—testing, rotation among departments, trial and error, etc.—for greater knowledge permits a more efficient utilization of manpower. Expenditures on acquiring knowledge of employee talents would be a specific investment if the knowledge could be kept from other firms, for then productivity would be raised more in the firms making the expenditures than elsewhere.

The effect of investment in employees on their productivity elsewhere depends on market conditions as well as on the nature of the investment. Very strong monopsonists might be completely insulated from competition by other firms, and practically all investments in their labor force would be specific. On the other hand, firms in extremely competitive labor markets would face a constant threat of raiding and would have fewer specific investments available.

These examples convey some of the surprisingly large variety of situations that come under the rubric of specific investment. This set is now treated abstractly in order that a general formal analysis can be developed. Empirical situations are brought in again after several major implications of the formal analysis have been developed.

If all training were completely specific, the wage that an employee could get elsewhere would be independent of the amount of training he had received. One might plausibly argue, then, that the wage paid by firms would also be independent of training. If so, firms would have to pay training costs, for no rational employee would pay for training that did not benefit him. Firms would collect the return from such training in the form of larger profits resulting from higher productivity, and training would be provided whenever the return—discounted at an appropriate rate—was at least as large as the cost. Long-run competitive equilibrium requires that the present value of the return exactly equals costs.

These propositions can be stated more formally with the equations developed earlier. According to equations (5) and (7) the equilibrium of a firm providing training in competitive markets can be written as

$$MP_0' + G\left[\sum_{t=1}^{n-1} \frac{MP_t - W_t}{(1+i)^t}\right] = W_0 + C, \tag{11}$$

where C is the cost of training given only in the initial period, MP_0' is the opportunity marginal product of trainees, W_0 is the wage paid to trainees, and W_t and MP_t are the wage and marginal product in period t. If the analysis of completely specific training given in the preceding paragraph was correct, W would always equal the wage that could

be received elsewhere, $MP_t - W_t$ would be the full return in t from training given in 0, and G would be the present value of these returns. Since MP_0' measures the marginal product elsewhere and W_0 would measure the wage elsewhere of trainees, $MP_0' = W_0$. As a consequence $G = C$, or, in full equilibrium, the return from training equals costs.

Before claiming that the usual equality between marginal product and wages holds when completely specific training is considered, the reader should bear in mind two points. The first is that the equality between wages and marginal product in the initial period involves opportunity, not actual marginal product. Wages would be greater than actual marginal product if some productivity was foregone as part of the training program. The second is that, even if wages equaled marginal product initially, they would be less in the future because the differences between future marginal products and wages constitute the return to training and are collected by the firm.

All of this follows from the assumption that firms pay all costs and collect all returns. But could not one equally well argue that workers pay all specific training costs by receiving appropriately lower wages initially and collect all returns by receiving wages equal to marginal product later? In terms of equation (11), W_t would equal MP_t, G would equal zero, and $W_0 = MP_0' - C$, just as with general training. Is it more plausible that firms rather than workers pay for and collect and return from training?

An answer can be found by reasoning along the following lines. If a firm had paid for the specific training of a worker who quit to take another job, its capital expenditure would be partly wasted, for no further return could be collected. Likewise, a worker fired after he had paid for specific training would be unable to collect any further return and would also suffer a capital loss. The willingness of workers or firms to pay for specific training should, therefore, closely depend on the likelihood of labor turnover.

To bring in turnover at this point may seem like a *deus ex machina* since it is almost always ignored in traditional theory. In the usual analysis of competitive firms, wages equal marginal product, and since wages and marginal product are assumed to be the same in many firms, no one suffers from turnover. It would not matter whether a firm's labor force always contained the same persons or a rapidly changing group. Any person leaving one firm could do equally well in other firms, and his employer could replace him without any change in profits. In other words, turnover is ignored in traditional theory because it plays no important role within the framework of the theory.

Turnover becomes important when costs are imposed on workers or firms, which are precisely the effects of specific training. Suppose a firm paid all the specific training costs of a worker who quit after completing it. According to our earlier analysis he would have been receiving the market wage and a new employee could be hired at the same wage. If the new employee were not given training, his marginal product would be less than that of the one who quit since presumably training raised the latter's productivity. Training could raise the new employee's productivity but would require additional expenditures by the firm. In other words, a firm is hurt by the departure of a trained employee because an equally profitable new employee could not be obtained. In the same way an employee who pays for specific train-

ing would suffer a loss from being laid off because he could not find an equally good job elsewhere. To bring turnover into the analysis of specific training is not, therefore, a *deus ex machina* but is made necessary by the important link between them.

Firms paying for specific training might take account of turnover merely by obtaining a sufficiently large return from those remaining to counterbalance the loss from those leaving. (The return on "successes"—those remaining—would, of course, overestimate the average return on all training expenditures.) Firms could do even better, however, by recognizing that the likelihood of a quit is not fixed but depends on wages. Instead of merely recouping on successes what is lost on failures, they might reduce the likelihood of failure itself by offering higher wages after training than could be received elsewhere. In effect, they would offer employees some of the return from training. Matters would be improved in some respects but worsened in others, for the higher wage would make the supply of trainees greater than the demand, and rationing would be required. The final step would be to shift some training costs as well as returns to employees, thereby bringing supply more in line with demand. When the final step is completed firms no longer pay all training costs nor do they collect all the return but they share both with employees.[14] The shares of each depend on the relation between quit rates and wages, layoff rates and profits, and on other factors not discussed here, such as the cost of funds, attitudes toward risk, and desires for liquidity.[15]

If training were not completely specific, productivity would increase in other firms as well, and the wage that could be received elsewhere would also increase. Such training can be looked upon as the sum of two components, one completely general, the other completely specific, with the former being relatively larger the greater the effect on wages in other firms relative to the firms providing the training. Since firms do not pay any of completely general costs and only part of completely specific costs, the fraction of costs paid by firms would be negatively related to the importance of the general component, or positively related to the specificity of the training.

Our conclusions can be stated formally in terms of the equations developed earlier. If G is the present value of the return from training collected by firms, the fundamental equation is

$$MP' + G = W + C . \qquad (12)$$

If G' measures the return collected by employees, the total return, G'', would be the sum of G and G'. In full equilibrium the total return would equal total costs, or $G'' = C$. Let a represent the fraction of the total return collected by firms. Since $G = aG''$ and $G'' = C$, equation (12) can be written as

[14] Marshall was clearly aware of specific talents and their effect on wages and productivity: "Thus the head clerk in a business has an acquaintance with men and things, the use of which he could in some cases sell at a high price to rival firms. But in other cases it is of a kind to be of no value save to the business in which he already is; and *then his departure would perhaps injure it by several times the value of his salary*, while probably he could not get half that *salary elsewhere*" (*op. cit.*, p. 626). (My italics.) However, he overstressed the element of indeterminacy in these wages ("their earnings are determined . . . by a bargain between them and their employers, the terms of which are theoretically arbitrary" [*ibid.*, fn.]) because he ignored the effect of wages on turnover.

[15] The rate used to discount costs and returns is the sum of a (positive) rate measuring the cost of funds, a (positive or negative) risk premium, and a liquidity premium that is presumably positive since capital invested in specific training is very illiquid (see the discussion in Sec. IV, *C*).

$$MP' + aC = W + C\,, \qquad (13)$$

or

$$W = MP' - (1 - a)C\,.^{16} \qquad (14)$$

Employees pay the same fraction of costs, $1 - a$, as they collect in returns, which generalizes the results obtained earlier. For if training were completely general, $a = o$, and equation (14) reduces to equation (10); if firms collected all the return from training, $a = 1$, and (14) reduces to $MP_0' = W_0$; if $0 < a < 1$, none of the earlier equations are satisfactory.

A few major implications of this analysis of specific training are now developed.

Rational firms pay generally trained employees the same wage and specifically trained employees a higher wage than they could get elsewhere. A reader might easily believe the contrary, namely, that general training would command a higher wage relative to alternatives than specific training does, since, after all, competition for persons with the latter is apt to be weaker than for those with the former. This view, however, overlooks the fact that general training raises the wages that could be received elsewhere while (completely) specific training does not, so a comparison with alternative wages gives a misleading impression of the *absolute* effect on wages of different types of training. Moreover, firms are not too concerned about the turnover of employees with general training and have no incentive to offer them a premium above wages elsewhere because the cost of such training is borne entirely by employees. Firms are concerned about the turnover of employees with specific training, and a premium is offered to reduce their turnover because firms pay part of their training costs.

The part of specific training paid by employees has effects similar to those discussed earlier for general training: it is also paid by a reduction in wages during the training period, tends to make age-earnings profiles steeper and more concave, etc. The part paid by firms has none of these implications, since current or future wages would not be affected.

Specific, unlike general, training would produce certain "external" effects, for quits would prevent firms from capturing the full return on costs paid by them, and layoffs would do the same to employees. Note, however, that these are external *diseconomies* imposed on the employees or employers of firms providing the training, not external economies accruing to other firms.

Employees with specific training have less incentive to quit, and firms have less incentive to fire them, than employees with no or general training, which implies that quit and layoff rates would be inversely related to the amount of specific training. Turnover would be least for employees with extremely specific training and most for those receiving such general training that productivity was raised less in firms providing the training than elsewhere. These propositions are as applicable to the large amount of irregular quits and layoffs that continually occur as to the more regular cyclical and secular movements in turnover; in this section, however, only the more regular movements are discussed.

Consider a firm that experiences an unexpected decline in demand for its

[16] If G'' did not equal C, these equations would be slightly more complicated. Suppose, for example, $G'' = G + G' = C + n$, $n \geq 0$, so that the present value of the total return would be greater than total costs. Then $G = aG'' = aC + an$, and

$$MP' + aC + an = W + C\,,$$

or

$$W = MP' - [(1 - a)C - an]\,.$$

output, the rest of the economy being unaffected. The marginal product of employees without specific training—such as untrained or generally trained employees—presumably initially equaled wages, and their employment would be reduced to prevent their marginal productivity from falling below wages. The marginal product of specifically trained employees initially would have been greater than wages. A decline in demand would reduce these marginal products too, but as long as they were reduced by less than the initial difference with wages, firms have no incentive to lay off such employees. For sunk costs are sunk, and there is no incentive to lay off employees whose marginal product is greater than wages, no matter how unwise it was, in retrospect, to invest in their training. Thus workers with specific training seem less likely to be laid off as a consequence of a decline in demand than are untrained or even generally trained workers.[17]

If the decline in demand were sufficiently great so that even the marginal product of specifically trained workers was pushed below wages, would the firm just proceed to lay them off until the marginal product was brought into equality with wages? To show the danger here, assume that all the cost and return from specific training was paid and collected by the firm. Any worker laid off would try to find a new job, since nothing would bind him to the old one.[18] The firm might be hurt if a new job was found, for the firm's investment in his training might be lost forever. If specifically trained workers were not laid off, the firm would lose now because marginal product would be less than wages but would gain in the future if the decline in demand proved temporary. There is an incentive, therefore, not to lay off workers with specific training when their marginal product is only temporarily below wages, and the larger a firm's investment the greater the incentive not to lay off such workers.

A worker collecting some of the return from specific training would have less incentive to find a new job when temporarily laid off than others would: he does not want to lose his investment. His behavior while laid off in turn affects his chances of being laid off, for if it were known that he would not readily take another job, the firm could lay him off without much fear of losing its investment.

The conclusion here can be briefly summarized. When one firm alone experiences an unexpected decline in demand, relatively few workers with specific training would be laid off, if only because their marginal product were initially greater than their wage. If the decline were permanent, all workers would be laid off when their marginal product became less than their wage and all those laid off would have to find jobs elsewhere. If the decline were temporary, specifically trained workers might not be laid off even though their marginal product were less than their wage because the firm would suffer if they took other jobs. The likelihood of their taking other jobs would be inversely related, and therefore the likelihood of their being laid off would be positively related, to the extent of their own investment in training.

The analysis can easily be extended to

[17] A very similar argument is developed by Walter Oi in "Labor as a Quasi-fixed Factor of Production" (unpublished Ph.D. dissertation, University of Chicago).

[18] Actually one need only assume that the quit rate of laid-off workers tends to be significantly greater than that of employed workers, if only because the cost of searching for another job is less for laid-off workers.

cover general declines in demand; suppose, for example, a general cyclical decline occurred. Let me assume that wages are sticky and remain at the initial level. If the decline in business activity were not sufficient to reduce the marginal product below the wage, workers with specific training would not be laid off even though others would be, just as before. If the decline reduced marginal product below wages, only one modification in the previous analysis is required. A firm would have a greater incentive to lay off specifically trained workers than when it alone experiences a decline because laid-off workers would be less likely to find other jobs when unemployment was widespread. In other respects the implications of a general decline with wage rigidity are the same as those of a decline in one firm alone.

The discussion has concentrated on layoff rates, but the same kind of reasoning shows that a rise in wages elsewhere would cause fewer quits among specifically trained workers than among others. For specifically trained workers initially receive higher wages than are available elsewhere and the wage rise elsewhere would have to be greater than the initial difference before they would consider quitting. Thus both the quit and layoff rate of specifically trained workers would be relatively low and fluctuate relatively less during business cycles. These are important implications than can be tested with the data available.

Although quits and layoffs are influenced by considerations other than investment costs, some of these, such as the presence of pension plans, are more strongly related to investments than may appear at first blush. A pension plan with incomplete vesting privileges[19] penalizes employees quitting before retirement and thus provides an incentive —often an extremely powerful one—not to quit. At the same time pension plans "insure" firms against quits for they are given a lump sum—the non-vested portion of payments—whenever a worker quits. Insurance is needed for specifically trained employees because their turnover would impose capital losses on firms. Firms can discourage such quits by sharing training costs and the return with employees, but they have less need to discourage them and would be more willing to pay for training costs if insurance was provided. The effects on the incentive to invest in one's employees may have been a major stimulus to the development of pension plans.[20]

An effective long-term contract would insure firms against quits, just as pensions do, and also insure employees against layoffs. Firms would be more willing to pay for all kinds of training—assuming future wages were set at an appropriate level—since a contract, in effect, converts all training into completely specific training. A casual reading of history suggests that long-term contracts have, indeed, primarily been a means of inducing firms to undertake large investments in employees. These contracts are seldom used today in the United States,[21] and while they have declined in importance over time, they were probably always the exception here largely because courts have considered them a form of involuntary servitude.

[19] According to the National Bureau of Economic Research study of pensions, most plans still have incomplete vesting (see D. Holland's report in *A Respect for Facts: National Bureau of Economic Research Annual Report* [New York: National Bureau of Economic Research, 1960], pp. 44–46).

[20] In recent years pensions have also been an important tax-saving device, which certainly has been a crucial factor in their mushrooming growth.

[21] The military and entertainment industry are the major exceptions.

Moreover, any enforcible contract could at best specify the hours required on a job, not the quality of performance. Since performance can vary widely, unhappy workers could usually "sabotage" operations to induce employers to release them from contracts.

Some training may be useful neither in most nor only in a single firm but in a set of firms defined by product, type of work, or geographical location. For example, carpentry training would raise productivity primarily in the construction industry, and French legal training would be ineffective in the United States, with its different language and legal institutions. Such training would tend to be paid by trainees, since a single firm could not readily collect the return,[22] and in this respect would be the same as general training. In one respect, however, it is similar to specific training. Workers with training "specific" to an industry, occupation, or country are less likely to leave that industry, occupation, or country (via migration) than other workers, so their industrial, occupational, or country "turnover" would be less than average. The same result is obtained for specific training, except that a firm rather than an industry, occupation, or country is used as the unit of observation in measuring turnover. An analysis of specific training, therefore, is helpful also in understanding the effects of certain types of "general" training.

Although a discrepancy between marginal product and wages is frequently taken as evidence of imperfections in the competitive system, it would occur even in a perfectly competitive environment where there is investment in specific training. The investment approach provides a very different interpretation of some common phenomena, as can be seen from the following examples.

A positive difference between marginal product and wages is usually said to be evidence of monopsony power, and just as the ratio of product price to marginal cost has been suggested as a measure of monopoly power, so has the ratio of marginal product to wages been suggested as a measure of monopsony power. But specific training would also make this ratio greater than one. Does the difference between the marginal product and the earnings of major-league baseball players, for example, measure monopsony power or the return on a team's investment? Since teams do spend a great deal on developing players, some and perhaps most of the difference must be considered a return on investment even were there no uncertainty about the abilities of different players.[23]

Earnings might differ greatly among firms, industries, and countries and yet there may be relatively little worker mobility. The usual explanation would be that workers were either irrational or faced with formidable obstacles in moving. However, if specific[24] training were important, differences in earnings would be a misleading estimate of what "migrants" could receive, and it might be perfectly rational not to move. For example, although French lawyers earn less than American lawyers, the average French lawyer could not earn the average American legal income simply by migrat-

[22] Sometimes firms co-operate in paying training costs, especially when training apprentices (see *A Look at Industrial Training in Mercer County, N.J.* [Washington Bureau of Apprenticeship and Training, 1959], p. 3).

[23] S. Rottenberg ("The Baseball Players' Labor Market," *Journal of Political Economy*, June, 1956, p. 254) argues that the strong restrictions on entry of teams into the major leagues is prima facie evidence that monopsony power is important, but the entry or threat of new *leagues*, such as have occurred in professional basketball and football, is a real possibility.

[24] Specific, that is, to the firms, industries, or countries in question.

ing to the United States, for he would have to invest in learning English and American law and procedures.[25]

In extreme types of monopsony, exemplified by an isolated company town, job alternatives for both trained and untrained workers are nil, and all training, no matter what the nature, would be specific to the firm. Monopsony combined with control of a product or an occupation (due, say, to anti-pirating agreements) converts training specific to that product or occupation into firm-specific training. These kinds of monopsony increase the importance of specific training and thus the inventive to invest in employees.[26] The effect on training of less extreme monopsony positions is more difficult to assess. Consider the monopsonist who pays his workers the best wage available elsewhere. I see no reason why training should have a systematically different effect on the foregone earnings of his employees than of those in competitive firms and, therefore, no reason why specific training should be more (or less) important to him. But monopsony power as a whole, including the more extreme manifestations, would appear to increase the importance of specific training and the incentive for firms to invest in human capital.

B. SCHOOLING

A school can be defined as an institution specializing in the production of training, as distinct from a firm that offers training in conjunction with the production of goods. Some schools, like those for barbers, specialize in one skill, while others, like universities, offer a large and diverse set. Schools and firms are often substitute sources of particular skills. The shift that has occurred over time in both law and engineering is a measure of this substitution. In acquiring legal skills the shift has been from apprenticeships in law firms to law schools, and in engineering skills from on-the-job experience to engineering schools.[27]

Some types of knowledge can be mastered better if simultaneously related to a practical problem; others require prolonged specialization. That is, there are complementarities between learning and work and between learning and time. Most training in the construction industry is apparently still best given on the job, while the training of physicists requires a long period of specialized effort. The development of certain skills requires both specialization and experience and can be had partly from firms and partly from schools. Physicians receive apprenticeship training as interns and residents after several years of concentrated instruction in medical schools. Or to take an example closer to home, a research economist not only spends many years in school but also a rather extensive apprenticeship in mastering the "art" of empirical and theoretical research. The complementarity with firms and schools depends in part on the amount of formalized knowledge available: price theory can be formally presented in a course, while a formal statement of the principles

[25] Of course, persons who have not yet invested in themselves would have an incentive to migrate, and this partly explains why young persons migrate more than older ones. For a further explanation see my discussion on p. 38; also see the paper in this Supplement by L. Sjaastad.

[26] A relatively large difference between marginal product and wages in monopsonies might measure, therefore, the combined effect of economic power and a relatively large investment in employees.

[27] State occupational licensing requirements often permit on-the-job training to be substituted for school training (see S. Rottenberg, "The Economics of Occupational Licensing" [paper given at the National Bureau of Economic Research Conference on Labor Economics, April, 1960]).

used in gathering and handling empirical materials is lacking.

Training in a new industrial skill is usually first given on the job, since firms tend to be the first to be aware of its value, but as demand develops, some of the training shifts to schools. For example, engineering skills were initially acquired on the job, and over time engineering schools have been developed.

A student does not work for pay while in school but may do so "after" or "before" school, or during "vacations." His earnings are usually less than if he were not in school since he cannot work as much or as regularly. The difference between what could have been and is earned is an important and indirect cost of schooling. Tuition, fees, books and supplies, unusual transportation and lodging expenses are other, more direct, costs. *Net* earnings can be defined as the difference between actual earnings and direct school costs. In symbols,

$$W = MP - k, \qquad (15)$$

where MP is actual marginal product (assumed equal to earnings) and k is direct costs. If MP_0 is the marginal product that could have been received, equation (15) can be written as

$$W = MP_0 - (MP_0 - MP + k) = MP_0 - C, \qquad (16)$$

where C is the sum of direct and foregone costs and where net earnings are the difference between potential earnings and total costs. These relations should be familiar since they are the same as those derived for general on-the-job training, which suggests that a sharp distinction between schools and firms is not always necessary: for some purposes schools can be treated as a special kind of firm and students as a special kind of trainee. Perhaps this is most apparent when a student works in an enterprise controlled by his school, which frequently occurs at many colleges.

Our definition of student net earnings may seem strange since tuition and other direct costs are not usually subtracted from "gross" earnings. Note, however, that indirect school costs are implicitly subtracted, for otherwise earnings would have to be defined as the sum of observed and foregoine earnings, and foregone earnings are a major cost of high school, college, and adult schooling. Moreover, earnings of on-the-job trainees would be net of *all* their costs, including direct "tuition" costs. Consistent accounting, which is particularly important when comparing earnings of persons trained in school and on the job, would require that earnings of students be defined in the same way.[28]

Regardless of whether all costs or merely indirect costs are subtracted from potential earnings, schooling would have the same kind of implications as general on-the-job training. Thus schooling would steepen the age-earnings profile, mix together the income and capital accounts, introduce a negative relative between the permanent and current earnings of young persons, and allow for depreciation on human capital. This supports our earlier assertion that an analysis of on-the-job training leads to general results that apply to other kinds of investment in human capital as well.

C. OTHER KNOWLEDGE

On-the-job and school training are not the only activities that raise real income primarily by increasing the knowledge at a person's command. Information

[28] Students often have negative net earnings and in this respect differ from most on-the-job trainees, although at one time many apprentices also had negative earnings.

about the prices charged by different sellers would enable a person to buy from the cheapest, thereby raising his command over resources, or information about the wages offered by different firms would enable him to work for the firm paying the highest (see Stigler's paper in this Supplement, pp. 94–105). In both examples information about the economic system, of consumption and production possibilities, is increased as distinct from knowledge of a particular skill. Information about the political or social system—the effect of different parties or social arrangements—could also significantly raise real incomes.[29]

Let us consider in more detail investment in information about employment opportunities. A better job might be found by spending money on employment agencies and situation-wanted ads, using one's time to examine want ads, talking to friends and visiting firms, or in Stigler's language by "search." When the new job requires geographical movement, additional time and resources would be spent in moving.[30] These expenditures constitute an investment in information about job opportunities that would yield a return in the form of higher earnings than would otherwise have been received. If workers paid costs and collected the return, an investment in search would have the same implications about age-earnings profiles, depreciation, and the like as general on-the-job training and schooling, although it must be noted that the direct costs of search, like the direct costs of schooling, are usually added to consumption rather than deducted from earnings. If firms paid costs and collected the return, search would have the same implications as on-the-job specific training.

Whether workers or firms pay for search depends on the effect of a job change on alternatives: the larger the number of alternatives made available by a change, the larger, not the smaller, the fraction of costs that have to be paid by workers. Consider a few examples. Immigrants to the United States usually found many firms that could use their talents, and these firms should have been reluctant to pay the large cost of transporting workers to the United States. In fact, immigrants almost always had to pay their own way. Even the system of contract labor, which we have seen is a means of protecting firms against turnover, was singularly unsuccessful in the United States and has been infrequently used.[31] Firms that are relatively insulated from competition in the labor market have an incentive to pay the costs of workers coming from elsewhere since they have little to worry about in the way of competing neighboring firms. In addition, firms would be willing partly to pay for search within a geographical area because some costs—such as an employment agency's fee—would be specific to the firm doing the hiring since they must be repeated at each job change.

D. PRODUCTIVE WAGE INCREASES

One way to invest in human capital is to improve emotional and physical health. In Western countries today earn-

[29] The role of political knowledge is systematically discussed in A. Downs, *An Economic Theory of Democracy* (New York: Harper & Bros., 1957), and more briefly in my "Competition and Democracy," *Journal of Law and Economics*, Vol. I (Fall, 1958).

[30] Studies of large geographical moves—those requiring both a change in employment and consumption—have tended to emphasize the job change more than the consumption change. Presumably money wages are considered to be more dispersed geographically than prices.

[31] For a careful discussion of the contract-labor system see C. Erickson, *American Industry and the European Immigrant, 1860–1885* (Cambridge, Mass.: Harvard University Press, 1957).

ings are much more closely geared to knowledge than to strength, but in an earlier day, and elsewhere still, strength had a significant influence on earnings. Moreover, emotional health increasingly is considered an important determinant of earnings in all parts of the world. Health, like knowledge, can be improved in many ways. A decline in the death rate at working ages may improve earning prospects by extending the period during which earnings are received; a better diet adds strength and stamina, and thus earning capacity; or an improvement in working conditions—higher wages, coffee breaks, and so on—might affect morale and productivity.

Firms can invest in the health of employees through medical examinations, luncheons, or steering them away from activities with high accident and death rates. An investment in health that increased productivity to the same extent in many firms would be a general investment and would have the same effect as general training, while an investment in health that increased productivity more in the firms making them would be a specific investment and would have the same effect as specific training. Of course, most investments in health in the United States are made outside firms, in households, hospitals, and medical offices. A full analysis of the effect on earnings of such "outside" investment in health is beyond the scope of this paper, but I would like to discuss a relation between on-the-job and "outside" human investments that has received much attention in recent years.

When on-the-job investments are paid by reducing earnings during the investment period, less is available for investments outside the job in health, better diet, schooling, and other factors. If these "outside" investments were more productive, some on-the-job investments would not be undertaken even though they were very productive by "absolute" standards.

Before I proceed further, one point needs to be made. The amount invested outside the job would be related to current earnings only if the capital market was very imperfect, for otherwise any amount of "outside" investment could be financed with borrowed funds. The analysis assumes, therefore, that the capital market is extremely imperfect, earnings and other income being a major source of funds.[32]

A firm would be willing to pay for investment in human capital made by employees outside the firm if it could benefit from the resulting increase in productivity. The only way to pay, however, would be to offer higher wages during the investment period than would have been offered since direct loans to employees are prohibited by assumption. When a firm gives a productive wage increase—that is, an increase that raises productivity—"outside" investments are, as it were, converted into on-the-job investments. Indeed, such a conversion is a natural way to circumvent imperfections in the capital market and the resultant dependence of the amount invested in human capital on the level of wages.

The discussion can be stated more formally. Let W represent wages in the absence of any investment, and let a productive wage increase costing an amount C be the only on-the-job investment. Total costs to the firm would be $\pi = W + C$, and since the investment cost is received by employees as higher wages, π would also measure total wages. The cost of on-the-job training is not

[32] Imperfections in the capital market with respect to investment in human capital are discussed in Sec. IV, *D*.

received as higher wages, so this formally distinguishes a productive wage increase from other on-the-job investments. The term MP can represent the marginal product of employees when wages equal W, and G the gain to firms from the investment in higher wages. In full equilibrium,

$$MP + G = W + C = \pi. \quad (17)$$

Investment would not occur if the firm's gain was nil ($G = o$), for then total wages (π) would equal the marginal product (MP) when there is no investment.

We have shown that firms would benefit more from on-the-job investment the more specific the productivity effect, the greater their monopsony power, and the longer the labor contract; conversely, the benefit would be less the more general the productivity effect, the less their monopsony power, and the shorter the labor contract. For example, a wage increase spent on a better diet with an immediate impact on productivity might well be granted,[33] but not one spent on general education with a very delayed impact.[34]

The effect of a wage increase on productivity depends on the way it is spent, which in turn depends on tastes, knowledge, and opportunities. Firms might exert an influence on spending by exhorting employees to consume good food, housing, and medical care, or even by requiring purchases of specified items in company stores. Indeed, the company store or truck system in nineteenth-century Great Britain has been interpreted as partly designed to prevent an excessive consumption of liquor and other debilitating commodities.[35] The prevalence of employer paternalism in underdeveloped countries has been frequently accepted as evidence of a difference in temperament between East and West. An alternative interpretation suggested by our study is that an increase in consumption has a greater effect on productivity in underdeveloped countries, and that a productivity advance raises profits more there either because firms have more monopsony power or because the advance is less delayed. In other words "paternalism" may simply be a way of investing in the health and welfare of employees in underdeveloped countries.

An investment in human capital would usually steepen age-earnings profiles, lowering reported earnings during the investment period and raising them later on. But an investment in an increase in earnings may have precisely the opposite effect, raising reported earnings more during the investment period than later and thus flattening age-earning

[33] The more rapid the impact the more likely that it comes within the (formal or de facto) contract period. Leibenstein apparently initially assumed a rapid impact when discussing wage increases in underdeveloped countries (see his "The Theory of Underemployment in Backward Economies," *Journal of Political Economy*, Vol. LXV [April, 1957]). In a later comment he argued that the impact might be delayed ("Underemployment in Backward Economies: Some Additional Notes," *Journal of Political Economy*, Vol. LXVI [June, 1958]).

[34] Marshall discusses delays of a generation or more and notes that profit-maximizing firms in competitive industries have no incentive to grant such wage increases.

"Again, in paying his workpeople high wages and in caring for their happiness and culture, the liberal employer confers benefits which do not end with his own generation. For the children of his workpeople share in them, and grow up stronger in body and in character than otherwise they would have done. The price which he has paid for labour will have borne the expenses of production of an increased supply of high industrial facilities in the next generation: but these facilities will be the property of others, who will have the right to hire them out for the best price they will fetch: neither he nor even his heirs can reckon on reaping much material reward for this part of the good that he has done" (*op. cit.*, p. 566).

[35] See G. W. Hilton, "The British Truck System in the Nineteenth Century," *Journal of Political Economy*, LXV (April, 1957), 246–47.

profiles. The cause of this difference is simply that reported earnings during the investment period tend to be net of the cost of general investments and gross of the cost of a productive earnings increase.[36]

The productivity of employees depends not only on their ability and the amount invested in them both on and off the job but also on their motivation, or the intensity of their work. Economists have long recognized that motivation in turn partly depends on earnings because of the effect of an increase in earnings on morale and aspirations. Equation (17), which was developed to show the effect of investments outside the firm financed by an increase in earnings, can also show the effect of an increase in the intensity of work "financed" by an increase in earnings. Thus W and MP would show initial earnings and productivity, C the increase in earnings, and G the gain to firms from the increase in productivity caused by the "morale" effect of the increase in earnings. The incentive to grant a morale-boosting increase in earnings, therefore, would depend on the same factors as does the incentive to grant an increase used for outside investments. Many recent discussions of wages in underdeveloped countries have stressed the latter,[37] while earlier discussions often stressed the former.[38]

[36] If E represents reported earnings during the investment period and MP the marginal product when there is no investment, $E = MP - C$ with a general investment, $E = MP$ with a specific investment paid by the firm, and $E = MP + C$ with a productive earnings increase.

[37] See the papers by Leibenstein, *op. cit.*, and H. Oshima, "Underdevelopment in Backward Economies: An Empirical Comment," *Journal of Political Economy*, Vol. LXVI (June, 1958).

[38] For example, Marshall stressed the effect of an increase in earnings on the character and habits of working people (*op. cit.*, pp. 529–32, 566–69).

III. RELATION BETWEEN EARNINGS, COSTS, AND RATES OF RETURN

Thus far little attention has been paid to the factors determining the amount invested in human capital. The most important single determinant is the profitability or rate of return, but the effect on earnings of a change in the rate of return has been difficult to distinguish empirically from a change in the amount invested. For investment in human capital usually extends over a long and variable period, so the amount invested cannot be determined from a known "investment period." Moreover, the discussion of on-the-job training clearly indicated that the amount invested is often merged with gross earnings into a single net earnings concept (which is gross earnings minus the cost or plus the return on investment).

In the following, some rather general relations between earnings, investment costs, and rates of return are derived. They permit one to distinguish, among other things, a change in the return from a change in the amount invested. The discussion proceeds in stages from simple to complicated situations. First, investment is restricted to a single period and returns to all remaining periods; then investment is permitted to be distributed over a known group of periods called the investment period. Finally, we show how the rate of return, amount invested, and the investment period can all be derived from information on net earnings alone.

Let Y be an activity providing a person entering at a particular age, called age zero, with a real net earnings stream of Y_0 during the first period, Y_1 the next period, and so on until Y_n is provided during the last period. The general term "activity" rather than occupation or

another more concrete term is used to indicate that any kind of investment in human capital is permitted, not just on-the-job training but also schooling, information, health, and morale. By "net" earnings I continue to mean that tuition costs during any period have been subtracted and returns added to "gross" earnings during the same period (see discussion in Sec. II). "Real" earnings are the sum of monetary earnings and the monetary equivalent of psychic earnings. Since many persons appear to believe that the term "investment in human capital" must be restricted to monetary costs and returns, let me emphasize that essentially all my analysis applies independently of the division of real earnings into monetary and psychic components. Thus the analysis applies to health, an activity with a large psychic component, as well as to on-the-job training, an activity with a large monetary component. When psychic components dominate, the language associated with consumer durable goods might be considered more appropriate than that associated with investment goods, but to simplify the presentation, I use investment language throughout.

The present value of the net earnings stream in Y would be

$$V(Y) = \sum_{j=0}^{n} \frac{Y_j}{(1+i)^{j+1}}, \quad (18)$$

[39]

where i is the market discount rate, assumed for simplicity to be the same in each period. If X were another activity providing a net earning stream of $X_0, X_1, \ldots X_n$, with a present value of $V(X)$, the present value of the gain from choosing Y would be given by

$$d = V(Y) - V(X) = \sum_{j=0}^{n} \frac{Y_j - X_j}{(1+i)^{j+1}}. \quad (19)$$

Equation (19) can be reformulated to bring out explicitly the relation between costs and returns. The cost of investing in human capital equals the net earnings foregone by choosing to invest rather than choosing an activity requiring no investment. If activity Y requires an investment only in the initial period and if X does not require any, the cost of choosing Y rather than X is simply the difference between their net earnings in the initial period, and the total return would be the present value of the differences between net earnings in later periods. If $C = X_0 - Y_0$, $k_j = Y_j - X_j$, $j = 1, \ldots n$, and if R measures the total return, the gain from Y could be written as

$$d = \sum_{j=1}^{n} \frac{k_j}{(1+i)^j} - C = R - C. \quad (20)$$

The relation between costs and returns can be derived in a different and, for our purposes, preferable way by defining the internal rate of return,[40] which is simply a rate of discount equating the present value of returns to the present value of costs. In other words, the internal rate, r, is defined implicitly by the equation

[39] Our discussion assumes discrete income flows and compounding, even though a mathematically more elegant formulation would have continuous variables, with sums replaced by integrals and discount rates by continuous compounding. The discrete approach is, however, easier to follow and yet yields the same kind of results as the continuous approach. Extensions to the continuous case are straightforward.

[40] A substantial literature has developed on the difference between the income gain and internal return approaches. See, for example, Friedrich and Vera Lutz, *The Theory of Investment of the Firm* (Princeton, N.J.: Princeton University Press, 1951), chap. ii, and the articles in *The Management of Corporate Capital*, ed. Ezra Solomon (Glencoe, Ill.: Free Press, 1959).

$$C = \sum_{1}^{n} \frac{k_j}{(1+r)^j}, \qquad (21)$$

which clearly implies

$$\sum_{j=0}^{n} \frac{Y_j}{(1+r)^{j+1}} - \sum_{0}^{n} \frac{X_j}{(1+r)^{j+1}} = d = 0, \qquad (22)$$

since $C = X_0 - Y_0$ and $k_j = Y_j - X_j$. So the internal rate is also a rate of discount equating the present values of net earnings. These equations would be considerably simplified if the return were the same in each period, or $Y_j = X_j + k$, $j = 1, \ldots n$. Thus equation (21) would become

$$C = \frac{k}{r}[1 - (1+r)^{-n}], \qquad (23)$$

where $(1+r)^{-n}$ is a correction for the finiteness of life that tends toward zero as people live longer.

If investment is restricted to a single known period, cost and rate of return are easily determined from information on net earnings alone. Since, however, investment in human capital is distributed over many periods—formal schooling is usually more than ten years in the United States, and long periods of on-the-job training are also common—the analysis must be generalized to cover distributed investment. The definition of an internal rate in terms of the present value of net earnings in different activities obviously applies regardless of the amount and duration of investment, but the definition in terms of costs and returns is not generalized so readily. If investment were known to occur in Y during each of the first m periods, a simple and superficially appealing approach would be to define the investment cost in each of these periods as the difference between net earnings in X and Y, total investment costs as the present value of these differences, and the internal rate would equate total costs and returns. In symbols,

$$C_j^1 = X_j - Y_j, \qquad j = 0, \ldots m-1,$$

$$C^1 = \sum_{0}^{m-1} C_j^1 (1+r)^{-j},$$

and

$$C^1 = \frac{k}{r}\left[\frac{1 - (1+r)^{m-1-n}}{(1+r)^{m-1}}\right]. \qquad (24)$$

If $m = 1$, this reduces to equation (23).

Two serious drawbacks mar this appealing straightforward approach. The estimate of total costs requires a priori knowledge and specification of the investment period. While the period covered by formal schooling is easily determined, the period covered by much on-the-job training and other investment is not, and a serious error might result from an incorrect specification: to take an extreme example, total costs would approach zero as the investment period is assumed to be longer and longer.[41]

A second difficulty is that the differences between net earnings in X and Y do not correctly measure the cost of investing in Y since they do not correctly measure earnings foregone. A person who invested in the initial period could receive more than X_1 in period 1 as long as the initial investment yielded a positive return.[42] The true cost of an invest-

[41] Since

$$C^1 = \sum_{0}^{m-1} (X_j - Y_j)(1+r)^{-j},$$

$$\lim_{m \to n} C^1 = \sum_{0}^{n-1} (X_j - Y_j)(1+r)^{-j} = 0,$$

by definition of the internal rate.

[42] If C_0 was the initial investment, r_0 its internal rate, and if the return were the same in all years, the amount

$$X_1^1 = X_1 + \frac{r_0 C_0}{1 - (1+r_0)^{-n}}$$

could be received in period 1.

ment in period 1 would be the total earnings foregone, or the difference between what could have been received and what is received. The difference between X_1 and Y_1 could greatly underestimate true costs; indeed, Y_1 might be greater than X_1 even though a large investment was made in period 1.[43] In general, therefore, the amount invested in any period would be determined not only from net earnings in the same period but also from net earnings in earlier periods.

If the cost of an investment is consistently defined as the earnings foregone, quite different estimates of total costs emerge. Although superficially a less natural and straightforward approach, the generalization from a single period to distributed investment is actually greatly simplified. So let C_j be the foregone earnings in the jth period, r_j the rate of return on C_j, and let the return per period on C_j be a constant k_j, with $k = \Sigma k_j$ being the total return on the whole investment. If the number of periods was indefinitely large, and if investment occurred only in the first m periods, the equation relating costs, returns, and internal rates has the strikingly simple form of[44]

$$C = \sum_0^{m-1} C_j = \frac{k}{\bar{r}}, \qquad (25)$$

where

$$\bar{r} = \sum_0^{m-1} w_j r_j, \qquad w_j = \frac{C_j}{C},$$

and

$$\sum_0^{m-1} w_j = 1. \qquad (26)$$

Total cost, defined simply as the sum of cost during each period, would equal the capitalized value of returns, the rate of capitalization being a weighted average of the rates of return on the individual investments. Any sequence of internal rates or investment costs is permitted, no matter what the pattern of rises and declines, nor what form the investments take, be they a college education, an apprenticeship, ballet lessons, or a medical examination. Different investment programs would have the same ultimate effect on earnings whenever the average rate of return and the sum of investment costs were the same.[45]

Equation (25) can be given an interesting interpretation if all rates of return were the same. The term k/r would then be the value at the beginning of the mth period of all succeeding net earning differentials between Y and X discounted

[43] Y_1 is greater than X_1 if

$$X_1 + \frac{r_0 C_0}{1 - (1 + r_0)^{-n}} - C_1 > X_1,$$

or if

$$\frac{r_0 C_0}{1 - (1 + r_0)^{-n}} > C_1,$$

where C_1 is the investment in period 1.

[44] A proof is straightforward. An investment in period j would yield a return of the amount $k_j = r_j C_j$ in each succeeding period if the number of periods was infinite and the return was the same in each. Since the total return is the sum of individual returns,

$$k = \sum_0^{m-1} k_j = \sum_0^{m-1} r_j C_j = C \sum_0^{m-1} \frac{r_j C_j}{C} = \bar{r} C.$$

I am indebted to Helen Raffel for important suggestions which led to this simple proof.

[45] Note that the rate of return equating the present values of net earnings in X and Y is not necessarily equal to $\bar{r}$, for it would weigh more heavily than $\bar{r}$ does the rates of return on earlier investments. For example, if rates were higher on investments in earlier than later periods, the over-all rate would be greater than $\bar{r}$, and vice versa if rates were higher in later periods. The difference between the over-all internal rate for X and Y and $\bar{r}$ would be small, however, as long as the investment period was not very long and the systematic difference between internal rates not very great.

at the internal rate, r.[46] Total costs would equal the value also at the beginning of the mth period—which is the end of the investment period—of the first m differentials between X and Y.[47] The value of the first m differentials between X and Y must equal the value of all succeeding differentials between Y and X, since r would be the rate of return equating the present values in X and Y.

The internal rate of return and the amount invested in each of the first m periods could be estimated from the net earnings streams in X and Y alone if the rate of return was the same on all investments. For the internal rate r could be determined from the condition that the present value of net earnings must be the same in X and Y, and the amount invested in each period seriatim from the relations[48]

$$C_0 = X_0 - Y_0, \quad C_1 = X_1 - Y_1 + rC_0$$
$$C_j = X_j - Y_j + r\sum_{k=0}^{j-1} C_k, 0 \leq j \leq m-1. \tag{27}$$

So costs and the rate of return can be estimated from information on net earnings. This is fortunate since the return on human capital is never empirically separated from other earnings and the cost of such capital is only sometimes and incompletely separated.

The investment period of education can be measured by years of schooling, but the period of on-the-job training, the search for information, and other investments is not readily available. Happily, one need not know the investment period to estimate costs and returns, since all three can be simultaneously estimated from information on net earnings. If activity X were known to have no investment (a zero investment period) the amount invested in Y during any period would be defined by

[46] That is,

$$\sum_{j=m}^{\infty} (Y_j - X_j)(1+r)^{m-1-j} = k\sum_{m}^{\infty}(1+r)^{m-1-j} = \frac{k}{r}.$$

[47] Since, by definition,

$$X_0 - Y_0 = C_0, \qquad X_1 - Y_1 = C_1 - rC_0,$$

and more generally

$$X_j - Y_j = C_j - r\sum_{k=0}^{j-1} C_k, \qquad 0 \leq j < m,$$

then

$$\sum_{j=0}^{m-1}(X_j - Y_j)(1+r)^{m-1-j} = \sum_{j=0}^{m-1}\left(C_j - r\sum_{0}^{j-1}C_k\right)(1+r)^{m-1-j}$$
$$= \sum_{0}^{m-1} C_j\{(1+r)^{m-1-j} - r[1 + (1+r) + \ldots + (1+r)^{m-2-j}]\}.$$
$$= \sum_{0}^{m-1} C_j = C.$$

The analytical difference between the naïve definition of costs advanced earlier and one in terms of foregone earnings is that the former measures total costs by the value of earning differentials at the beginning of the investment period and the latter by the value at the end of the period. Therefore, $C^1 = C(1+r)^{1-m}$, which follows from eq. (24) when $n = \infty$.

[48] If the rate of return was not the same on all investments there would be $2m$ unknowns—$C_0, \ldots C_m - 1$, and $r_0, \ldots r_{m-1}$—and only $m+1$ equations—the m cost definitions and the equation

$$k = \sum_{0}^{m-1} r_i C_i.$$

An additional $m-1$ relation would be required to determine the $2m$ unknowns. The condition $r_0 = r_1 = \ldots = r_{m-1}$ is one form these $m-1$ relations can take.

$$C_j = X_j - Y_j + r \sum_0^{j-1} C_k, \quad \text{all } j, \quad (28)$$

and total costs by

$$C = \sum_0^{\infty} C_j. \quad (29)$$

The internal rate could be determined in the usual way from the equality between present values in X and Y, costs in each period from equation (28) and total costs from equation (29).

The definition of costs presented here simply extends to all periods the definition advanced earlier for the investment period.[49] The rationale for the general definition is the same: investment occurs in Y whenever earnings there are below the sum of those in X and the income accruing on prior investments. If costs were found to be greater than zero before some period m and equal to zero thereafter, the first m periods would be the empirically derived investment period. But costs and returns can be estimated from equation (28) even when there is no simple investment period.

A common objection to an earlier draft of this paper is that the general and rather formal definition of costs advanced here is all right when applied to on-the-job training, schooling, and other recognized investments, but goes too far by also including as investment costs many effects that should be treated otherwise. For example, the protest runs, suppose that learning was essentially unavoidable in an activity Z, so that earnings "automatically" grows rapidly with experience. Since earnings in Z would tend to be lower than those in X at younger ages and higher later on, my approach would say that investment occurs in Z. Critics have argued that there really is no investment in Z since the rise in earnings results from *unavoidable* learning rather than from an attempt to improve skills, knowledge, or health. Although the argument is superficially plausible I am convinced it is as reasonable to say that investment in human capital occurs in Z as in activities requiring training or schooling. Indeed, an important virtue rather than defect in my concept of human capital is that learning—both on and off the job—is included along with training and schooling.

If Z were preferred to X the higher earnings at later ages presumably outweigh the earnings foregone initially. Similarly, a person entering an activity requiring much education is said to value

[49] Therefore, since the value of the first m earning differentials has been shown to equal

$$\sum_0^{m-1} C_j$$

at period m (see n. 47), total costs could be estimated from the value of all differentials at the end of the earning period. That is,

$$C = \sum_0^{\infty} C_j = \sum_0^{\infty} (X_j - Y_j)^{\infty-1-j}.$$

Thus the value of all differentials would equal zero at the beginning of the earning period—by definition of the internal rate—and C at the end. The apparent paradox results from the infinite horizon, as can be seen from the following equation relating the value of the first f differentials at the beginning of the gth period to costs:

$$V(f, g) = \sum_{j=0}^{f-1} (X_j - Y_j)(1+r)^{g-1-j}$$

$$= \sum_{j=0}^{f-1} C_j (1+r)^{g-f}.$$

When $f = \infty$ and $g = 0$, $V = 0$, but whenever $f = g$,

$$V = \sum_0^{f-1} C_j.$$

In particular, if $f = g = \infty$, $V = C$.

the stream of future higher earnings more than the net earnings foregone initially. If the lower earnings due to education are called investment costs, the higher earnings investment returns, and if costs are related to returns by an internal rate of return, logical consistency and economic sense would require that similar concepts apply to learning. Thus the lower initial earnings of high-school graduates who enter occupations "with a future" have as much right to be considered investment, both from the social and private viewpoints, as do the lower net earnings of those enrolled in college. In general, since the private and social ranking of different economic activities depend only on their net earning streams, if one activity was said to require a given investment and to yield a given return, another activity with the same net earning stream must be said to require the same investment and yield the same return, no matter how they differ in other respects.

So much in defense of our approach. To estimate costs empirically still has required a priori knowledge that nothing is invested in activity X. Without such knowledge, only the *difference* between the amounts invested in any two activities with known net earning streams could be estimated from the definitions in equation (28). Were this done for all available streams the investment in any activity beyond that in the activity with the smallest investment could be determined.[50] The observed minimum investment would not be zero, however, if the rate of return on some initial investment was sufficiently high to attract everyone. A relevant question is, therefore: can the shape of the stream in an activity having zero investment be specified a priori so that the total investment in any activity can be determined?

The statement "nothing is invested in an activity" means only nothing would be invested after the age when information on earnings first became available; investment can have occurred before that age. If, for example, the data begin at age eighteen, some investment in schooling, health, or information surely must have occurred at younger ages. The earning stream of persons who do not invest after age eighteen would have to be considered, at least in part, as a return on the investment before eighteen. Indeed, in the developmental approach to child-rearing (discussed in Selma Mushkin's paper), most if not all of these earnings would be so considered.

The earning stream in an activity with no investment beyond the initial age (activity X) would be flat if the developmental approach was followed and earnings were said to result entirely from earlier investment.[51] The minimum investment could then be determined if an assumption was made about its rate to return. My discussion of the shape of the earning stream in X is, however, highly conjectural,[52] and further investigation may well indicate that another approach is preferable.

Our assumption that lifetimes are infinite, although descriptively unrealistic, is often a very close approximation. For example, I have shown elsewhere that the average rate of return on college education in the United States could

[50] The technique is applied and further developed by Mincer in his paper in this Supplement.

[51] If C measured the cost of investment before the initial age and r its rate of return, $k = rC$ would measure the return per period. If earnings were attributed entirely to this investment, $X_i = k = rC$, where X_i represents earnings at the ith period past the initial age.

[52] But note that empirical evidence indicates that age-earning profiles in unskilled occupations are very flat.

only be slightly raised if people remained in the labor force indefinitely. A finite earning period has, however, a greater effect on the rate of return of investments occurring at later ages, say after age forty; indeed, it helps explain why schooling and other investments are primarily made at younger ages.

An analysis of finite earning streams can be approached in two ways. One simply applies the concepts developed for infinite streams and says there is disinvestment in human capital when net earnings are above the amount that could be maintained indefinitely. Investment at younger ages would give way to disinvestment at older ages until no human capital remained at death (or retirement). This approach has several important applications and is used in parts of my study. An alternative that is more useful for some purposes lets the earning period itself influence the definitions of accrued income and cost. The income resulting from an investment during period j would be defined as

$$k_j = \frac{r_j C_j}{1 - (1 + r_j)^{j-n}}, \qquad (30)$$

where $n + 1$ is the earning period, and the amount invested during j would be defined by

$$C_j = X_j - Y_j + \sum_{k=0}^{k=j-1} \frac{r_k C_k}{1 - (1 + r_k)^{k-n}}. \qquad (31)$$

IV. THE INCENTIVE TO INVEST

A. NUMBER OF PERIODS

The discussion summarized in equations (28) and (31) shows how total costs, rates of return, and the investment period can be estimated from information on net earnings alone, and thus how the effect on earnings of a change in the amount invested can be distinguished empirically from the effect of a change in rates of return. Our attention now turns to the factors influencing the amount invested in different activities and by different persons. Economists have long believed that the incentive to expand and improve physical resources depends on the rate of return expected. They have been very reluctant, however, to interpret improvements in the effectiveness and amount of human resources in the same way, namely, as systematic responses or "investments" resulting in good part from the returns expected. In this section I try to show that an investment approach to human resources is a powerful and simple tool capable of explaining a wide range of phenomena, including much that has either been ignored or given *ad hoc* interpretations.

An increase in the lifespan of an activity would, other things the same, increase the rate of return on the investment made in any period. The influence of lifespan on the rate of return and thus on the incentive to invest is important and takes many forms. A few of these forms will now be discussed.

The number of periods is obviously affected by mortality and morbidity rates, for the lower they are, the longer the expected lifespan, and the larger the fraction of a lifetime that can be spent at any activity. The major secular decline of these rates in the United States and elsewhere may have increased the rates of return on investment in human capital,[53] thereby encouraging such investment. This conclusion is independent

[53] I say *may* because rates of return are adversely affected by the increase in labor force that would result from a decline in death and sickness. If the adverse effect was sufficiently great, a decline in death and sickness would reduce rates of return on human capital. I am indebted to my wife for emphasizing this point.

of whether the secular improvement in health itself resulted from investment; if so, the secular increase in rates of return would be part of the return to investment in health.

A relatively large fraction of younger persons are in school, enter upon on-the-job training, change jobs and locations, and add to their knowledge of economic, political, and social opportunities. The entire explanation of these differences between young and old persons may not be that the young are more interested in learning, more able to absorb new ideas, less tied down by family responsibilities, more easily supported by parents, or more flexible about changing their routine and place of living. One need not rely only on life-cycle effects on capabilities, responsibilities, or attitudes as soon as one recognizes, as we have throughout, that schooling, training, mobility, and the like are ways to invest in human capital and that younger people have a greater incentive to invest because they can collect the return over more years.[54] Indeed, a greater incentive would be present even if age had no effect on capabilities, responsibilities, and attitudes.

Although the unification of these different kinds of behavior by the investment approach is important evidence in its favor, other evidence is needed. A powerful test can be developed along the following lines.[55] Suppose that investment in human capital raised earnings for p periods only, where p varied between o and n. The size of p would be affected by many factors, including the rate of obsolescence since the more rapidly an investment became obsolete the smaller p would be. The advantage in being young would be less the smaller p was, since the effect of age on the rate of return would be positively related to p. For example, if p equaled two years, the rate would be the same at all ages except the two nearest the "retirement" age. If the investment approach was correct, the difference between the amount invested at different ages would be positively correlated with p, which is not surprising since an expenditure with a small p would be less of an "investment" than one with a large p, and arguments based on an investment framework would be less applicable. None of the life-cycle arguments seem to imply any correlation with p, so this provides a powerful test of the importance of the investment approach.

The time spent in any one activity is determined not only by age, mortality, and morbidity but also by the amount of switching between activities. Women spend less time in the labor force than men and, therefore, have less incentive to invest in market skills; tourists spend little time in any one area and have less incentive than residents of the area to invest in knowledge of specific consumption opportunities;[56] temporary migrants to urban areas have less incentive to invest in urban skills than permanent residents; and, as a final example, draftees have less incentive than professional soldiers to invest in purely military skills.

Women, tourists, and the like have to

[54] Younger persons would also have a greater incentive to invest if the cost of any investment rose with age, say, because potential and thus foregone earnings rose with age.

[55] This test was suggested by George Stigler's discussion of the effect of different auto-correlation patterns on the incentive to invest in information (see "The Economics of Information," *Journal of Political Economy*, Vol. LXIX [June, 1961], and his paper in this Supplement).

[56] This example is from Stigler, "The Economics of Information," *op. cit.*

find investments that increase productivity in several activities. A woman wants her investment to be useful both as a housewife and as a participant in the labor force, or a frequent traveler wants to be knowledgeable in many environments. Such investments would be less readily available than more specialized ones—after all, an investment increasing productivity in two activities also increases it in either one alone, extreme complementarity aside, while the converse does not hold; specialists, therefore, have greater incentive to invest in themselves than others do.

Specialization in an activity would be discouraged if the market were very limited; thus the incentive to specialize and to invest in oneself would increase as the extent of the market increased. Workers would be more skilled the larger the market, not only because "practice makes perfect," so often stressed in discussions of the division of labor,[57] but also because a larger market would *induce* a greater investment in skills.[58] Put differently, the usual analysis of the division of labor stresses that efficiency, and thus wage rates, would be greater the larger the market, and ignores the potential earnings period in any activity, while ours stresses that this period, and thus the incentive to *become* more efficient, would be directly related to market size. Surprisingly little attention has been paid to the influence of market size on the incentive to invest in skills.

[57] See, for example, Marshall, *op. cit.*, Bk. IV, chap. ix.

[58] If "practice makes perfect" means that age-earnings profiles slope upward, then according to my approach it must be treated along with other kinds of learning as a way of investing in human capital. The distinction above between the effect of an increase in the market on practice and on the incentive to invest would simply be that the incentive to invest in human capital is increased even aside from the effect of practice on earnings.

B. WAGE DIFFERENTIALS AND SECULAR CHANGES

According to equation (30) the internal rate of return depends on the ratio of the return per unit time to investment costs. A change in the return and costs by the same percentage would not change the internal rate, while a greater percentage change in the return would change the internal rate in the same direction. The return is measured by the absolute income gain, or by the absolute income difference between persons differing only in the amount of their investment. Note that absolute, not relative, income differences determine the return and the internal rate.

Occupational and educational wage differentials are sometimes measured by relative, sometimes by absolute, wage differences,[59] although no one has adequately discussed their relative merits. Marginal productivity analysis relates the derived demand for any class of workers to the ratio of their wages to those of other inputs,[60] so wage ratios are more appropriate in understanding forces determining demand. They are not, however, the best measure of forces determining supply, for the return on investment in skills and other knowledge is determined by absolute wage differences.

[59] See A. M. Ross and W. Goldner, "Forces Affecting the Inter-industry Wage Structure," *Quarterly Journal of Economics*, Vol. LXIV (May, 1950); P. H. Bell, "Cyclical Variation and Trend in Occupational Wage Differentials in American Industry since 1914," *Review of Economics and Statistics*, Vol. XXIII (November, 1951); F. Meyers and R. L. Bowlby, "The Interindustry Wage Structure and Productivity," *Industrial and Labor Relations Review*, Vol. VII (October, 1953); Stigler and Blank, *op. cit.*, Table 11; P. Keat, "Long-Term Trends in Occupational Wage Differentials," *Journal of Political Economy*, Vol. LXVIII (December, 1960).

[60] Thus the elasticity of a substitution is usually defined as the percentage change in the ratio of quantities employed per 1 per cent change in the ratio of wages.

Therefore neither wage ratios nor wage differences are uniformly the best measure, ratios being more appropriate in demand studies and differences in supply studies.

The importance of distinguishing between wage ratios and differences, and the confusion resulting from the practice of using ratios to measure supply as well as demand forces, can be illustrated by considering the effects of technological progress. If progress were uniform in all industries and neutral with respect to all factors, and if there were constant costs, initially all wages would rise by the same proportion and the prices of all goods, including the output of industries supplying the investment in human capital,[61] would be unchanged. Since wage ratios would be unchanged, firms would have no incentive initially to alter their factor proportions. Wage differences, on the other hand, would rise at the same rate as wages, and since investment costs would be unchanged, there would be an incentive to invest more in human capital, and thus to increase the relative supply of skilled persons. The increased supply would in turn reduce the rate of increase of wage differences and produce an absolute narrowing of wage ratios.

In the United States during much of the last eighty years, a narrowing of wage ratios has gone hand in hand with an increasing relative supply of skill, an association that is usually said to result from the effect of an *autonomous* increase in the supply of skills—brought about by the spread of free education or the rise in incomes—on the return to skill, as measured by wage ratios. An alternative interpretation suggested by our analysis is that the spread of education and the increased investment in other kinds of human capital were in large part *induced* by technological progress (and perhaps other changes) through the effect on the rate of return, as measured by wage diffrences and costs. Clearly a secular decline in wage ratios is not inconsistent with a secular increase in real wage differences if average wages were rising, and, indeed, one important body of data on wages shows a decline in ratios and an even stronger rise in differences.[62]

The interpretation based on autonomous supply shifts has been favored partly because a decline in wage ratios has erroneously been taken as evidence of a decline in the return to skill. While a decision ultimately can be based only on a detailed re-examination of the evidence,[63] the induced approach can be made more plausible by considering trends in physical capital. Economists have been aware that the rate of return on capital could be rising or at least not falling while the ratio of the "rental" price of capital to wages was falling. Consequently, although the rental price

[61] Some persons have argued that only direct investment costs would be unchanged, indirect costs or foregone earnings rising along with wages. Neutral progress implies, however, the same increase in the productivity of a student's time as in his teacher's time or in the use of raw materials, so even foregone earnings would not change.

[62] Keat's data for 1906–53 in the United States show both an average annual decline of 0.8 per cent in the coefficient of variation of wages and an average annual rise of 1.2 per cent in the real standard deviation. The decline in the coefficient of variation was shown in his study (*op. cit*); I computed the change in the real standard deviation from data made available to me by Keat.

[63] For those believing that the evidence overwhelmingly indicates a secular decline in rates of return on human capital, I reproduce Adam Smith's statement on earnings in some professions. "The lottery of the law, therefore, is very far from being a perfectly fair lottery; and that, as well as many other liberal and honourable professions, is, in point of pecuniary gain, evidently under-recompensed" (*The Wealth of Nations* [New York: Modern Library, 1937], p. 106). Since economists tend to believe that law and most other liberal professions are now overcompensated relative to non-professional work "in point of pecuniary gain," the return to professional work could not have declined continuously if Smith's observations were accurate.

of capital declined relative to wages over time, the large secular increase in the amount of physical capital per man-hour is not usually considered autonomous, but rather induced by technological and other developments that, at least temporarily, raised the return. A common explanation based on the effects of economic progress may, then, account for the increase in both human and physical capital.

C. RISK AND LIQUIDITY

An informed, rational person would invest only if the expected rate of return was greater than the sum of the interest rate on riskless assets and the liquidity and risk premiums associated with the investment. Not much need be said about the "pure" interest rate, but a few words are in order on risk and liquidity. Since human capital is a very illiquid asset—it cannot be sold and is rather poor collateral on loans—a positive liquidity premium, perhaps a sizable one, would be associated with such capital.

The actual return on human capital varies around the expected return because of uncertainty about several factors. There always has been considerable uncertainty about the length of life, one important determinant of the return. People are also uncertain about their ability, especially younger persons who do most of the investing. In addition, there is uncertainty about the return to a person of given age and ability because of numerous events that are not predictable. The long time required to collect the return on an investment in human capital reduces the knowledge available, for required is knowledge about the environment when the return is to be received, and the longer the average period between investment and return the less such knowledge is available.

Informed observation as well as calculations I have made suggest that there is much uncertainty about the return to human capital.[64] The response to uncertainty is determined by its amount and nature and by tastes or attitudes. Many have argued that attitudes of investors in human capital are very different from those of investors in physical capital because the former tend to be younger,[65] and young persons are supposed to be especially prone to overestimate their ability and chance of good fortune.[66] Were this view correct, a human investment which promised a large return to exceptionally able or lucky persons would be more attractive than a similar physical investment. However, a "life-cycle" explanation of attitudes toward risk may be no more valid or necessary than life-cycle explanations of why investors in human capital are relatively young (discussed on pp. 37–38). Indeed, an alternative explanation of reactions to large gains has already appeared.[67]

[64] For example, Marshall said: "Not much less than a generation elapses between the choice by parents of a skilled trade for one of their children, and his reaping the full results of their choice. And meanwhile the character of the trade may have been almost revolutionized by changes, on which some probably threw long shadows before them, but others were such as could not have been foreseen even by the shrewdest persons and those best acquainted with the circumstances of the trade" (*op. cit.*, p. 571), and "the circumstances by which the earnings are determined are less capable of being foreseen [than those for machinery]" (*ibid.*).

[65] Note that our argument on p. 38 implied that investors in human capital would be younger.

[66] Smith said: "The contempt of risk and the presumptuous hope of success, are in no period of life more active than at the age at which young people choose their professions" (*op. cit.*, p. 109). Marshall said that "young men of an adventurous disposition are more attracted by the prospects of a great success than they are deterred by the fear of failure" (*op. cit.*, p. 554).

[67] See M. Friedman and L. J. Savage, "The Utility Analysis of Choices Involving Risk," reprinted in *Readings in Price Theory*, ed. G. J. Stigler and K. Boulding (Chicago: Richard D. Irwin, Inc., 1952).

D. CAPITAL MARKETS AND KNOWLEDGE

If investment decisions respond only to earning prospects, adjusted for risk and liquidity, the adjusted marginal rate of return would be the same on all investments. The rate of return on education, training, migration, health, and other human capital is supposed to be higher than elsewhere, however, because of financing difficulties and inadequate knowledge of opportunities. These will now be discussed briefly.

Economists have long emphasized that it is difficult to borrow funds to invest in human capital because such capital cannot be offered as collateral and courts have frowned on contracts which even indirectly suggest involuntary servitude. This argument has been explicitly used to explain the "apparent" underinvestment in education and training and also, although somewhat less explicitly, underinvestment in health, migration, and other human capital. The importance attached to capital market difficulties can be determined not only from the discussions of investment but also from the discussions of consumption. Young persons would consume relatively little, productivity and wages might be related, and some other consumption patterns would follow only if it were difficult to capitalize future earning power. Indeed, unless capital limitations applied to consumption as well as investment, the latter could be indirectly financed with "consumption" loans.[68]

Some other implications of capital market difficulties can also be mentioned:

1. Since large expenditures would be more difficult to finance, investment in (say) a college education would be more affected than in (say) short-term migration.

2. Internal financing would be common, and consequently wealthier families would tend to invest more than poorer ones.

3. Since employees' specific skills are part of the intangible assets or good will of firms and can be offered as collateral along with tangible assets, capital would be more readily available for specific than for general investments.

4. Some persons have argued that opportunity costs (foregone earnings) are more readily financed than direct costs because they require only to do "without," while the latter require outlays. Although superficially plausible, this view can easily be shown to be wrong: opportunity and direct costs can be financed equally readily, given the state of the capital market. If total investment costs were \$800, potential earnings \$1,000, and if all costs were foregone earnings, investors would have \$200 of earnings to spend; if all were direct costs, they would initially have \$1,000 to spend, but just \$200 would remain after paying "tuition," so their *net* position would be exactly the same as before. The example can be readily generalized and the obvious inference is that indirect and direct investment costs are equivalent in imperfect as well as perfect capital markets.

While it is undeniably difficult to use the capital market to finance investments in human capital, there is some reason to doubt whether otherwise equivalent investments in physical capital can be financed much more easily. Consider an eighteen-year-old who wants to invest a given amount in equipment

[68] A person with an income of X and investment costs of Y ($Y < X$) could either use X for consumption and receive an *investment loan* of Y, or use $X - Y$ for consumption, Y for investment, and receive a *consumption loan* of Y. He ends up with the same consumption and investment in both cases, the only difference being in the names attached to loans.

for a firm he is starting rather than in a college education. What is his chance of borrowing the whole amount at a "moderate" interest rate? Very slight, I believe, since he would be untried and have a high debt equity ratio; moreover, the collateral provided by his equipment would probably be very imperfect. He, too, would either have to borrow at high interest rates or self-finance. Although the difficulties of financing investments in human capital have usually been related to special properties of human capital, in large measure they seem also to beset comparable investments in physical capital.

A recurring theme is that young persons are especially prone to be ignorant of their abilities and of the investment opportunities available. If so, investors in human capital, being younger, would be less aware of opportunities and thus more likely to err than investors in tangible capital. I suggested earlier (pp. 37–38) that investors in human capital are younger partly because of the cost in postponing their investment to older ages. The desire to acquire additional knowledge about the return and about alternatives provides an incentive to postpone any risky investment, but since an investment in human capital is more costly to postpone, it would be made earlier and presumably with less knowledge than comparable non-human investments. Therefore, investors in human capital may not have less knowledge *because* of their age; rather both might be a *joint* product of the incentive not to delay investing.[69]

[69] Marshall (*op. cit.*, pp. 571–73) appears to argue that it is also intrinsically more difficult to acquire knowledge about the return from an investment in human capital.

The eighteen-year-old in our example who could not finance a purchase of machinery might, without too much cost, postpone the investment for a number of years until his reputation and equity were sufficient to provide the "personal" collateral required to borrow funds. Financing may prove a more formidable obstacle to investors in human capital because they cannot postpone their investment so readily. Perhaps this accounts for the tendency of economists to stress capital market imperfections when discussing investments in human capital.

V. SOME EFFECTS OF HUMAN CAPITAL

A. EXAMPLES

Differences in earnings among persons, areas, or time periods are usually said to result from differences in physical capital, technological knowledge, ability, or institutions (such as unionization or socialized production). Our analysis indicates, however, that investment in human capital also has an important effect on observed earnings because earnings tend to be net of investment costs and gross of investment returns. Indeed, an appreciation of the direct and indirect importance of human capital appears to resolve many otherwise puzzling empirical findings about earnings. Consider the following examples:

1. Almost all studies show that age-earnings profiles tend to be steeper among more skilled and educated persons. I argued earlier (pp. 14–15) that on-the-job training would steepen age-earning profiles and the analysis of Section III generalizes the argument to all human capital. Since observed earnings are gross of returns and net of costs, investment in human capital at younger ages would reduce observed earnings then and raise them at older ages, thus steepening

the age-earnings profile.[70]

2. In recent years students of international trade theory have been somewhat shaken by findings that the United States, said to have relative scarcity of labor and abundance of capital, apparently exports relatively labor-intensive commodities and imports relatively capital-intensive commodities. For example, one study found that export industries pay higher wages than import competing ones.[71]

An interpretation consistent with the Ohlin-Heckscher emphasis on the relative abundance of different factors argues that the United States has an even more (relatively) abundant supply of human than of physical capital. An increase in human capital would, however, show up as an apparent increase in labor intensity since earnings are gross of the return on such capital. Thus export industries might pay higher wages than import competing ones primarily because they employ more skilled or healthier workers.[72]

3. Several recent studies have tried to estimate empirically the elasticity of substitution between capital and labor. Usually a ratio of the input of physical capital to the input of labor is regressed on the wage rate in different areas or time periods, the regression coefficient being an estimate of the elasticity of substitution.[73] Countries, states, or time periods that have relatively high wages and inputs of physical capital also tend to have much human capital. Just as a correlation between wages, physical capital and human capital seems to obscure the relationship between relative factor supplies and commodity prices, so it obscures the relationship between relative factor supplies and factor prices. For if wages were high primarily because of human capital, a regression of the relative amount of physical capital on wages could give a seriously biased picture of the effect of factor proportions on wages.[74]

[70] According to eq. (28) earnings at age j can be approximated by

$$Y_j = X_j + \sum_{k=0}^{k=j-1} r_k C_k - C_j,$$

where X_j are earnings at j of persons who have not invested in themselves, C_k is the investment at age k, and r_k is its rate of return. The rate of increase in earnings would be at least as steep in Y as in X at each age and not only from "younger" to "older" ages if and only if

$$\frac{\Delta Y_j}{\Delta j} \geq \frac{\Delta X_j}{\Delta j},$$

or

$$r_j C_j \geq \frac{\Delta C_j}{\Delta j}.$$

This condition is usually satisfied since $r_j C_j \geq 0$ and the amount invested tends to decline with age.

[71] See I. Kravis, "Wages and Foreign Trade," *Review of Economics and Statistics*, Vol. XXXIII (February, 1956).

[72] This kind of interpretation has been put forward by many writers; see, for example, the discussion in W. Leontief, "Factor Proportions and the Structure of American Trade: Further Theoretical and Empirical Analysis," *Review of Economics and Statistics*, Vol. XXXIII (November, 1956).

[73] Interstate estimates for several industries can be found in J. Minasian, "Elasticities of Substitution and Constant-Output Demand Curves for Labor," *Journal of Political Economy*, LXIX (June, 1961), 261–70; intercountry estimates in Kenneth Arrow, Hollis B. Chenery, Bagicha Minhas, and Robert M. Solow, "Capital-Labor Substitution and Economic Efficiency," *Review of Economics and Statistics* (August, 1961); unpublished papers by Philip Nelson and Robert Solow contain both interstate and time-series estimates.

[74] Minasian's argument (*op. cit.*, p. 264) that interstate variations in skill level necessarily bias his estimates toward unity is actually correct only if skill is a perfect substitute for "labor." (In correspondence Minasian states that he intended to make this condition explicit.) If, on the other hand, human and physical capital were perfect substitutes the estimates would always have a downward bias, regardless of the true substitution between labor and capital. Perhaps the most reasonable assumption would be that physical capital is more complementary with human capital than with labor; I have not, however, been able to determine the direction of bias in this case.

4. A secular increase in average earnings has usually been said to result from increases in technological knowledge and physical capital per earner. The average earner, in effect, is supposed to benefit indirectly from activities by entrepreneurs, investors, and others. Another explanation put forward in recent years argues that earnings can rise because of direct investment in earners.[75] Instead of only benefiting from activities by others, the average earner is made a prime mover of development through the investment in himself.[76]

B. ABILITY AND THE DISTRIBUTION OF EARNINGS

An emphasis on human capital not only helps explain differences in earnings over time and among areas but also among persons or families within an area. This application will be discussed in greater detail than the others because a link is provided among earnings, ability, and the incentive to invest in human capital.

Economists have long been aware that conventional measures of ability—intelligence tests or aptitude scores, school grades, and personality tests—while undoubtedly relevant at times, do not reliably measure the talents required to succeed in the economic sphere. The latter requires a particular kind of personality, persistence, and intelligence. Accordingly, some writers have gone to the opposite extreme and argued that the only relevant way to measure economic talent is by results, or by earnings themselves.[77] Persons with higher earnings would simply have more ability than others, and a skewed distribution of earnings would imply a skewed distribution of abilities. This approach goes too far, however, in the opposite direction. The main reason for an interest in relating ability to earning is to distinguish its effects from differences in education, training, health, and other such factors, and a definition equating ability and earnings *ipso facto* precludes such a distinction. Nevertheless, results are very relevant and should not be ignored.

A compromise might be reached through defining ability by earnings only when several variables had been held constant. Since the public is very concerned about separating ability from education, on-the-job training, health, and other human capital, the amount invested in such capital would have to be held constant. Although a full analysis would also hold discrimination, nepotism, and several other factors constant, a reasonable first approximation would say that if two persons have the same investment in human capital, the one who earns more is demonstrating greater economic talent.

Since observed earnings are gross of the return on human capital they are affected by changes in the amount and rate of return. Indeed, after the investment period earnings (Y) can be simply approximated by

$$Y = X + rC\,, \qquad (32)$$

[75] The major figure here undoubtedly is T. W. Schultz. Of his many articles see esp. "Education and Economic Growth" in *Social Forces Influencing American Education* (Sixtieth Yearbook of the National Society for the Study of Education, Part II [Chicago: University of Chicago Press, 1961]).

[76] One caveat is called for, however. Since observed earnings are not only gross of the return from investments in human capital but also are net of some costs, an increased investment in human capital would both raise and reduce earnings. Although average earnings would tend to increase as long as the rate of return was positive, the increase is less than it would be if the cost of human capital, like that of physical capital, was not deducted from national income.

[77] Let me state again that whenever the word "earnings" appears I mean real earnings, or the sum of monetary earnings and the monetary equivalent of psychic earnings.

where C measures total investment costs, r the average rate of return, and X earnings when there is no investment in human capital. If the distribution of X is ignored for now, Y would depend only on r when C was held constant, so "ability" would be measured by the average rate of return on human capital.[78]

The amount invested is not the same for everyone, nor even in a very imperfect capital market rigidly fixed for any given person, but depends in part on the rate of return. Persons receiving a high marginal rate of return would have an incentive to invest more than others.[79] Since marginal and average rates are presumably positively correlated[80] and since ability is measured by the average rate, one can say that abler persons would invest more than others. The end result would be a positive correlation between ability and the investment in human capital,[81] a correlation with several important implications.

[78] Since r is a function of C, Y would indirectly as well as directly depend on C, and therefore the distribution of ability would depend on the amount of human capital. Some persons might rank high in earnings and thus high in ability if everyone were unskilled, and quite low if education and other training were widespread.

[79] In addition, they would find it easier to invest if the marginal return and the resources of parents and other relatives were positively correlated.

[80] According to a well-known formula

$$r_m = r_a\left(1+\frac{1}{e_a}\right),$$

where r_m is the marginal rate of return, r_a the average rate, and e_a the elasticity of the average rate with respect to the amount invested. The rates r_m and r_a would be positively correlated unless r_a and $1/e_a$ were sufficiently negatively correlated.

[81] This kind of argument is not new; Marshall argued that business ability and the ownership of physical capital would be positively correlated: "[economic] forces . . . bring about the result that there is a far more close correspondence between the ability of business men and the size of the businesses which they own than at first sight would appear probable" (*op. cit.*, p. 312).

One is that the tendency for abler persons to migrate, continue their education,[82] and generally invest more in themselves can be explained without recourse to an assumption that non-economic forces or demand conditions favor them at higher investment levels. A second implication is that the separation of "nature from nurture" or ability from education and other environmental factors is apt to be difficult, for high earnings would tend to signify both more ability and a better environment. Thus the earnings differential between college and high-school graduates does not measure the effect of college alone since college graduates are abler and would earn more even without the additional education. Or reliable estimates of the income elasticity of demand for children have been difficult to obtain because higher income families also invest more in contraceptive knowledge.[83]

The main implication, however, is in the field of personal income distribution. At least ever since the time of Pigou economists have tried to reconcile the strong skewness in the distribution of earnings and other income with a presumed symmetrical distribution of abilities.[84] Pigou's own solution, that property income is not symmetrically distributed, does not directly help explain the skewness in earnings. Subsequent attempts have largely concentrated on developing *ad hoc* random and other probabilistic mechanisms that have little

[82] The first is frequently alleged (see, for example, Marshall, *op. cit.*, pp. 199, 684). Evidence on the second is discussed in my forthcoming study for the National Bureau of Economic Research.

[83] See my "An Economic Analysis of Fertility" in *Demographic and Economic Change in Developed Countries* (Princeton, N.J.: Princeton University Press, 1960).

[84] See A. C. Pigou, *The Economics of Welfare* (4th ed.; London: Macmillan & Co., 1950), Part IV, chap. ii.

relation to the mainstream of economic thought.[85] The approach presented here, however, offers an explanation that is not only consistent with economic analysis but actually relies on one of its fundamental tenets; namely, that the amount invested is a function of the rate of return expected. In conjunction with the effect of human capital on earnings this tenet can explain several well-known properties of earnings distributions.

By definition, the distribution of earnings would be exactly the same as the distribution of ability if everyone invested the same amount in human capital; in particular, if ability were symmetrically distributed, earnings would also be. Equation (32) shows that the distribution of earnings would be exactly the same as the distribution of investment if all persons were equally able; again, if investment were symmetrically distributed, earnings would also be.[86] If ability and investment both varied, earnings would tend to be skewed even when ability and investment were not, but the skewness would be small as long as the amount invested was statistically independent of ability.[87]

Our analysis has shown, however, that abler persons would tend to invest more than others, so ability and investment would be positively correlated, perhaps quite strongly. Now the product of two symmetrical distributions is more positively skewed the higher the positive correlation between them, and might be quite skewed.[88] The economic incentive given abler persons to invest relatively large amounts in themselves does seem capable, therefore, of reconciling a strong positive skewness in earnings with a presumed symmetrical distribution of abilities.

Variations in X help explain an important difference among skill categories in the degree of skewness. The smaller the fraction of total earnings resulting

[85] A sophisticated example can be found in B. Mandelbrot, "The Pareto-Levy Law and the Distribution of Income," *International Economic Review*, Vol. I (May, 1960). In a recent paper, however, Mandelbrot has brought in maximizing behavior (see "Paretian Distributions and Income Maximization," *Quarterly Journal of Economics*, Vol. LXXVI [February, 1962]).

[86] Jacob Mincer ("Investment in Human Capital and Personal Income Distribution," *Journal of Political Economy*, Vol. LXVI [August, 1958]) concluded that a symmetrical distribution of investment in education implies a skewed distribution of earnings because he defines educational investment by school years rather than costs. If we follow Mincer in assuming that everyone was equally able, that schooling was the only investment, and that the cost of the nth year of schooling equaled the earnings of persons with $n-1$ years of schooling, then, say, a normal distribution of schooling can be shown to imply a log-normal distribution of school costs, and thus a log-normal distribution of earnings.

The difference between the earnings of persons with $n-1$ and n years of schooling would be $k_n = Y_n - Y_{n-1} = r_n C_n$. Since r_n is assumed to equal r for all n, and $C_n = Y_{n-1}$, this equation becomes $Y_n = (1+r)Y_{n-1}$, and therefore

$$C_1 = Y_0$$

$$C_2 = Y_1 = Y_0(1+r)$$

$$C_3 = Y_2 = Y_1(1+r) = Y_0(1+r)^2$$

$$C_n = Y_{n-1} = \ldots = Y_0(1+r)^{n-1},$$

or the cost of each additional year of schooling increases at a constant *rate*. Since total costs have the same distribution as $(1+r)^n$, a symmetrical, say a normal, distribution of school years, n, implies a log-normal distribution of costs and hence by eq. (32) a log-normal distribution of earnings. I am indebted to Mincer for a helpful discussion of the comparison and especially for the stimulation provided by his pioneering work. Incidentally, his article and the dissertation on which it is based cover a much broader area than has been indicated here.

[87] For example, C. C. Craig has shown that the product of two independent normal distributions is only slightly skewed (see his "On the Frequency Function of XY," *Annals of Mathematical Statistics*, VII [March, 1936], 3).

[88] Craig (*op. cit.*, pp. 9–10) showed that the product of two normal distributions would be more positively skewed the higher the positive correlation between them, and that the skewness would be considerable with high correlations.

from investment in human capital—the smaller rC relative to X—the more would the distribution of earnings be dominated by the distribution of X. Higher skill categories have a greater average investment in human capital and thus presumably a large rC relative to X. The distribution of "unskilled ability," X, would, therefore, tend to dominate the distribution of earnings in relatively unskilled categories while the distribution of a product of ability and the amount invested, rC, would dominate in skilled categories. Hence if abilities were symmetrically distributed, earnings would tend to be more symmetrically distributed among the unskilled than among the skilled.[89]

Equation (32) holds only when investment costs are small, which tends to be true at later ages, say after age thirty-five. Net earnings ar earlier ages would be given by

$$Y_j = X_j + \sum_{0}^{j-1} r_i C_i + (-C_j),$$

where j refers to the current year and i to previous years, C_i measures the investment cost of age i, C_j current costs, and r_i the rate of return on C_i. The distribution of $-C_j$ would be an important determinant of the distribution of Y_j since investment is large at these ages. Hence our analysis would predict a smaller (positive) skewness at younger than at older ages because the presumed negative correlation between $-C_j$ and $\sum_{0}^{j-1} r_i C_i$ would counteract the positive correlation between ability and investment.

A simple analysis of the incentive to invest in human capital seems capable of explaining, therefore, not only why the over-all distribution of earnings is more skewed than the distribution of abilities, but also why earnings are more skewed among older and skilled persons than among younger and less skilled ones. The renewed interest in investment in human capital may provide the means of bringing the theory of personal income distribution back into economics.

VI. SUMMARY AND CONCLUSIONS

Most investments in human capital both raise observed earnings at older ages, because returns are added to earnings then, and lower them at younger ages, because costs are deducted from earnings then. Since these common effects are produced by very different kinds of human capital, a basis is provided for a unified and powerful theory. The analysis proceeded from a discussion of specific kinds of human capital, with greatest attention paid to on-the-job training because it clearly illustrates and emphasizes the common effects, to a general theory applying to any kind.

The general theory has a wide variety of important implications, ranging from interpersonal and interarea differences in earnings, to the shape of age-earning profiles, to the effect of specialization on skill. For example, since earnings are gross of the return on human capital, some persons may earn more than others simply because they invest more in themselves. And since "abler" persons tend to invest more than others, the distribution

[89] As noted earlier, X does not really represent earnings when there is no investment in human capital, but only earnings when there is no investment after the initial age (be it fourteen, twenty-five, or six). Indeed, the developmental approach to child-rearing argues that earnings would be close to zero if there was no investment at all in human capital. The distribution of X, therefore, would be at least partly determined by the distribution of investment before the initial age, and if it and ability were positively correlated, X might be positively skewed, even though ability was not.

of earnings could be very unequal and even skewed, even though "ability" were symmetrically and not too unequally distributed. To take another example, learning, both on and off the job, and other activities appear to have exactly the same effects on observed earnings as do education, training, and other traditional investments in human capital. We argue that a relevant concept should cover all activities with identical effects and show that the total amount invested in a generalized concept of human capital and its rate of return can be estimated from information on earnings alone.

Some investments in human capital do not affect earnings because costs are paid and returns are collected by the firms, industries, or countries using the capital. These "specific" investments range from hiring costs to executive training and are more important than is commonly believed. To take a couple of examples, we showed that the well-known greater unemployment among unskilled than skilled workers may result from the latter having more specific capital; or incompletely vested pension plans may be a means of insuring firms against a loss on their specific investments.

This paper has concentrated on developing a theory of investment in human capital, with an emphasis on empirical implications rather than on formal generalization. Of course, empirical usefulness is the only justification for any theory, and although I did not try to bring in even the quite limited evidence on the role of human capital, the empirical work reported in this volume, my own work, and that of many others support the view that investment in human capital is a pervasive phenomenon and a valuable concept. The next few years should provide much stronger evidence on whether the recent emphasis placed on this concept is just another fad or a development of great and lasting importance.