Volume Title: Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education, Second Edition

Volume Author/Editor: Gary S. Becker

Volume Publisher: NBER

Volume ISBN: 0-226-04109-3

Volume URL: http://www.nber.org/books/beck75-1

Publication Date: 1975

Chapter Title: Investment in Human Capital: Rates of Return

Chapter Author: Gary S. Becker

Chapter URL: http://www.nber.org/chapters/c3734

Chapter pages in book: (p. 45 - 144)

CHAPTER III

Investment in Human Capital: Rates of Return

The most important single determinant of the amount invested in human capital may well be its 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 since investment in human capital usually extends over a long and variable period, 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 of or plus the return on investment).

1. Relation between Earnings, Costs, and Rates of Return

In this section, some important 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 distributed over a known group of periods called the investment period. Finally, it is shown how the rate of return, the amount invested, and the investment period can all be derived from information on net earnings alone.

The discussion is from the viewpoint of workers and is, therefore, restricted to general investments; since the analysis of specific investments and firms is very similar, its discussion is omitted.

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 during the next period, and so on until Y_n during the last period. The general term "activity" rather than occupation or another more concrete term is used in order 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. As in the previous chapter, "net" earnings mean "gross" earnings during any period minus tuition costs during the same period. "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 the whole analysis applies independently of the division of real earnings into monetary and psychic components. Thus the analysis applies to health, which has a large psychic component, as well as to on-the-job training, which has 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; to simplify the presentation, investment language is used throughout.

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

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

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, . . . X_n, with a present value of $V(X)$, the present value of the gain from choosing Y would be given by

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

[1] The 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 yields the same kind of results. Extensions to the continuous case are straightforward.

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

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

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,[2] 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

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

which clearly implies

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

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

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

[2] 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, 1951, Chapter ii, and the articles in *The Management of Corporate Capital,* Ezra Solomon, ed., Glencoe, 1959.

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 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, however, 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,

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

$$C^1 = \sum_0^{m-1} C_j^1(1 + r)^{-j},$$

and

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

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.[3]

[3] Since

$$C^1 = \sum_0^{m-1} (X_j - Y_j)(1 + r)^{-j}, \quad \lim_{m \to n} C^1 = \sum_0^{n-1} (X_j - Y_j)(1 + r)^{-j} = 0,$$

by definition of the internal rate.

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.[4] The true cost of an investment 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.[5] 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. Therefore, let C_j be the foregone earnings in the j^{th} 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 were indefinitely large, and if investment occurred only in the first m periods, the equation relating costs, returns, and internal rates would have the strikingly simple form of[6]

[4] 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.

[5] Y_1 is greater than X_1 if

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

where C_1 is the investment in period 1.

[6] 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 were infinite and the return were the same in each. Since the total return is the sum of individual returns,

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

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

$$C = \sum_{0}^{m-1} C_j = \frac{k}{\bar{r}}, \tag{25}$$

where

$$\bar{r} = \sum_{0}^{m-1} w_j r_j, \quad w_j = \frac{C_j}{C},$$

and

$$\sum_{0}^{m-1} w_j = 1. \tag{26}$$

Total cost, defined simply as the sum of costs 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, or the form of investments, 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.[7]

Equation (25) could 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 m^{th} period of all succeeding net earning differentials between Y and X discounted at the internal rate, r.[8] Total costs would equal the value also at the beginning of the m^{th} period—which is the end of the investment period—of the first m differentials between X and Y.[9] The value of the first m differentials between X

[7] 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 weight the rates of return on earlier investments more heavily than $\bar{r}$ does. For example, if rates were higher on investments in earlier than in later periods, the overall rate would be greater than $\bar{r}$, and vice versa if rates were higher in later periods. Sample calculations indicate, however, that the difference between the overall rate and $\bar{r}$ tends to be small as long as the investment period was not very long and the systematic difference between internal rates not very great.

[8] That is,

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

[9] Since, by definition,

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

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 were 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[10]

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

$$C_j = X_j - Y_j + r\sum_{k=0}^{j-1} C_k, \quad 0 \leq j \leq m - 1.^{11} \tag{27}$$

---

and more generally

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

then

$$\sum_{j=0}^{m-1} (X_j - Y_j)(1+r)^{m-1-j} = \sum_{j=0}^{m-1} \left(C_j - r\sum_{0}^{j-1} C_k\right)(1+r)^{m-1-j}$$

$$= \sum_{0}^{m-1} C_j\{(1+r)^{m-1-j} - r[1 + (1+r) + \cdots + (1+r)^{m-2-j}]\}$$

$$= \sum_{0}^{m-1} C_j = C.$$

The analytical difference between the naive 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 equation (24) when $n = \infty$.

[10] If the rate of return were 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 only one form these $m - 1$ relations can take; another is that costs decrease at certain known rates. If the latter were assumed, all the r_i could be determined from the earnings data.

[11] In econometric terminology this set of equations forms a "causal chain" because of the natural time ordering provided by the aging process. Consequently, there is no identification or "simultaneity" problem.

Thus 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 periods of on-the-job training, of the search for information, and of other investments are 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

$$C_j = X_j - Y_j + r\sum_{0}^{j-1} C_k, \quad \text{all } j, \tag{28}$$

and total costs by

$$C = \sum_{0}^{\infty} C_j. \tag{29}$$

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.[12] The

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

$$\sum_{0}^{m-1} C_i$$

at period m (see footnote 9), 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 g^{th} 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 $f = g = \infty$, $V = C$.

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 was 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.

Thus, so the protest might run, learning would automatically lead to a convex and relatively steep earnings profile not because of any associated investment in education or training, but because the well-known "learning curve" is usually convex and rather steep. Since the method presented here, however, depends only on the shape of age-earnings profiles, the effect of learning would be considered an effect of investment in human capital. I accept the argument fully; indeed, I believe that it points up the power rather than the weakness of my analysis and the implied concept of human capital.

To see this requires a fuller analysis of the effect of learning. Assume that Z permits learning and that another activity X does not and has a flat earnings profile: Z might have the profile labeled TT in Chart 1 (in Chapter II) and X that labeled UU. If TT were everywhere above UU—i.e., earnings in Z were greater than those in X at each age—there would be a clear incentive for some persons to leave X and enter Z. The result would be a lowering of TT and raising of UU; generally the process would continue until TT was no longer everywhere above UU, as in Chart 1. Earnings would now be lower in Z than in X at younger ages and higher only later on, and workers would have to decide whether the later higher earnings compensated for the lower initial earnings.

They presumably would decide by comparing the present value of earnings in Z and X, or, what is equivalent, by comparing the rate of return that equates these present values with rates that could be obtained elsewhere. They would choose Z if the present value were greater there, or if the equalizing rate were greater than those elsewhere. Therefore, they would choose Z only if the rate of return on their learning were sufficiently great, that is, only if the returns from learning—the higher earnings later on—offset the costs of learning—the lower earnings initially. Thus choosing between activities "with a future" and "dead-end" activities involves exactly the same considerations as choosing between continuing one's education and entering the labor force—whether returns in the form of higher subsequent earnings sufficiently offset costs in the form of lower initial ones. Although learning cannot be avoided once in activities like *Z*, it can be avoided beforehand because workers can enter activities like *X* that provide little or no learning. They or society would choose learning only if it were a sufficiently good investment in the same way that they or society would choose on-the-job training if it were sufficiently profitable.

Consequently, the conclusion must be that learning is a way to invest in human capital that is formally no different from education, on-the-job training, or other recognized investments. So it is a virtue rather than a defect of our formulation of costs and returns that learning is treated symmetrically with other investments. And there is no conflict between interpretations of the shape of earning profiles based on learning theory[13] and those based on investment in human capital because the former is a special case of the latter. Of course, the fact that the physical and psychological factors associated with learning theory[14] are capable of producing rather steep concave profiles, like *TT* and even *T′T′* in Chart 1, should make one hesitate in relating them to education and other conventional investments. The converse is also true, however: the fact that many investments in human capital in a market economy would produce "the learning curve" should make one hesitate in relating it to the various factors associated with learning theory.

Another frequent criticism is that many on-the-job investments are really free in that earnings are not reduced at any age. Although this would be formally consistent with my analysis since the rate of return need only be considered infinite (in Chart 1, *TT* would be nowhere below *UU*), I suspect that a closer examination of the alleged "facts" would usually reveal a much more conventional situation. For example, if abler employees were put through executive training programs, as is probable, they might earn no less than employees outside the programs but they might earn less than if they had not been in training.[15] Again, the earnings of employees receiving specific training may

[13] See, for example, J. Mincer, "Investment in Human Capital and Personal Income Distribution," *Journal of Political Economy*, August 1958, pp. 287–288.

[14] See, for example, R. Bush and F. Mosteller, *Stochastic Models for Learning*, New York, 1955.

[15] Some indirect evidence is cited by J. Mincer in "On-the-Job Training: Costs, Returns, and Some Implications," *Investment in Human Beings*, NBER Special Conference 15, supplement to *Journal of Political Economy*, October 1962, p. 53.

not be reduced for the reasons presented in Chapter II. Finally, one must have a very poor opinion of the ability of firms to look out for their own interests to believe that infinite rates of return are of great importance.

So much in defense of the approach. To estimate costs empirically still requires 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.[16] The observed minimum investment would not be zero, however, if the rate of return on some initial investment were sufficiently high to attract everyone. A relevant question is, therefore: can the shape of the stream in an activity with 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" only means that nothing was 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 eighteeen. Indeed, in the developmental approach to child rearing, 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 were followed and earnings were said to result entirely from earlier investment.[17] The incorporation of learning into the concept of investment in human capital also suggests that earnings profiles would be flat were there no (additional) investment. Finally, the empirical evidence, for what it is worth (see comments in Chapter VII), suggests that earnings profiles in unskilled occupations are quite flat. If the earnings profile in X were flat, the unobserved investment could easily be determined in the usual way once an assumption were made about its rate of return.

[16] The technique has been applied and developed further by Mincer (*ibid.*).

[17] 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.

The assumption that lifetimes are infinite, although descriptively unrealistic, often yields results that are a close approximation to the truth. For example, I show later (see Chapter VI, section 2) that the average rate of return on college education in the United States would be only 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 made at later ages, say, after 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 the study (see especially Chapter VII). 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

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

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

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

Addendum: The Allocation of Time and Goods over Time

Basic Model

This section discusses the allocation of time and goods over a lifetime among three main sectors: consumption, investment in human capital, and labor force participation. It uses the framework developed in my "A Theory of the Allocation of Time," *Economic Journal,* September 1965. That paper, however, considered the allocation only at a moment of time among various kinds of consumption and time utilizations; this discussion generalizes the analysis to decisions over time and to investment in human capital.

Assume that a person is certain that he will live n periods. His economic welfare depends on his consumption over time of objects of choice called commodities, as in

$$U = U(C_i, \ldots C_n), \tag{32}$$

where C_i is the amount of the commodity consumed during period i. As assumed in the paper cited above, C_i is in turn produced "at home" with inputs of his market goods and his own time. Let the (composite) market goods used in period i be x_i, and the (composite) amount of time combined with x_i be t_{c_i}. Then

$$C_i = {}_if(x_i, t_{ci}), \quad i = 1, \ldots n \tag{33}$$

where ${}_if$ is the production function in period i. If initially it is assumed that time can be allocated only between consumption and labor force participation (called "work"), the following identity holds in each period

$$t_{c_i} + t_{w_i} = t, \quad i = 1, \ldots n \tag{34}$$

where t_{w_i} is the amount of work in i, and t, the total time available during i, is independent of i if all periods are equally long.

The "endowment" in each period is not simply a fixed amount of "income" since that is affected by the hours spent at work, which is a decision variable. Instead it is the vector (w_i, v_i), where v_i is the amount of property income and w_i is the wage rate available in the i^{th} period.

Suppose that there is a perfect capital market with an interest rate, r, the same in each period. Then a constraint on goods that complements the constraints on time given by (34) is that the present value of expenditures on goods must equal the present value of incomes:[18]

$$\sum_{i=1}^{n} \frac{p_i x_i}{(1+r)^{i-1}} = \sum_{i=1}^{n} \frac{w_i t_{w_i} + v_i}{(1+r)^{i-1}}. \tag{35}$$

[18] Savings in period i is defined as

$$S_i = w_i t_{w_i} + v_i - p_i x_i.$$

Our formulation is implicitly assuming that the savings process itself takes no time; a somewhat weaker assumption, say that savings is less time-intensive than consumption, would not result in greatly different conclusions. I. Ehrlich and U. Ben-Zion have since analyzed the effect of time on savings in "A Theory of Productive Savings," University of Chicago, 1972.

Substitution for t_{w_i} from equation (34) into (35) gives the set of constraints

$$\sum_{i=1}^{n} \frac{p_i x_i + w_i t_{c_i}}{(1+r)^{i-1}} = \sum_{i=1}^{n} \frac{w_i t + v_i}{(1+r)^{i-1}}, \tag{36}$$

and

$$0 \leq t_{c_i} \leq t, \quad x_i \geq 0. \qquad i = 1, \ldots n \tag{37}$$

The term on the right equals "full wealth," which is an extension of the definition of "full income" given in my earlier article. The term on the left shows how this full wealth is "spent": either on goods or on the foregone earnings associated with the use of time in consumption. Each person (or family) is assumed to maximize his utility function given by equation (32) subject to the constraints given by (36) and (37), and the production functions given by (33). The decision variables are the t_{c_i} and x_i, $2n$ variables. If the optimal values of these variables are assumed to be in the interior of the regions given by (37), and if the wage rates w_i are independent of x_i and t_{c_i}, the first order optimality conditions are simply

$$U_i\, {}_if_x = \frac{\lambda p_i}{(1+r)^{i-1}} \qquad i = 1, \ldots n \tag{38}$$

$$U_i\, {}_if_t = \frac{\lambda w_i}{(1+r)^{i-1}} \qquad i = 1, \ldots n \tag{39}$$

where

$${}_if_x = \frac{\partial\, {}_if}{\partial x_i}, \quad {}_if_t = \frac{\partial\, {}_if}{\partial t_{c_i}}, \quad U_i = \frac{\partial U}{\partial C_i}$$

and λ is a Lagrangian multiplier equal to the marginal utility of wealth.

Dividing equation (39) by (38) gives

$$\frac{{}_if_t}{{}_if_x} = \frac{w_i}{p_i}, \qquad i = 1, \ldots n \tag{40}$$

or in each period the marginal product of consumption time relative to goods equals the real wage rate in the same period, and is independent of the interest rate. In other words, consumption time should have a relatively high marginal product when the real wage rate is relatively high.

To understand the implications of equation (40) somewhat better, assume that all ${}_if$ are homogeneous of the first degree, which is a fairly innocuous assumption in the present context. Let us also temporarily assume that the productivity of goods and consumption time do not vary with age, so that f's are the same. Since the marginal productivities of linear homogeneous production functions depend only on factor proportions, equation (40) implies, if marginal products are declining, with these additional assumptions that the production of commodities is relatively time-intensive when real wages are relatively low, and relatively goods-intensive when real wages are relatively high.

Note that this last result is a "substitution" effect and is unambiguous: it is not offset by any "income" effect that operates in the opposite direction. There is no offsetting income or wealth effect because "full" wealth is *fixed*, by the right-hand side of equation (36), and is *completely* independent of the allocation of time and goods over time or at a moment in time. Note, however, that this "substitution" effect is in terms of the *relative* time or goods intensities in different periods, and *not* in terms of the *absolute* amount of consumption time (sometimes called "leisure") in different periods. The latter cannot be determined from equation (40) alone, and depends on the allocation of commodities over time. Only if the consumption of commodities were the same at all periods would relative and absolute intensities necessarily move in the same direction.

To see what happens to commodity consumption over time, consider an alternative form of equation (38):

$$\frac{U_i}{U_j} = \frac{p_i f_{x_i}}{p_j f_{x_i}}(1+r)^{(j-i)}. \qquad i, j = 1, \dots n \quad (41)$$

If prices are assumed to be stable, $p_i = p_j = 1$ and equation (41) becomes

$$\frac{U_i}{U_j} = \frac{f_{x_j}}{f_{x_i}}(1+r)^{(j-i)}. \quad (42)$$

It has been shown that $t_{c_i}/x_i > t_{c_j}/x_j$ if $w_j > w_i$. It follows from the assumptions of homogeneity and diminishing returns that $f_{x_i} > f_{x_j}$. Hence from (42)

$$\frac{U_i}{U_j} \lesseqgtr (1+r)^{(j-i)} \quad \text{as} \quad w_i \lesseqgtr w_j. \quad (43)$$

Note that equality of the left- and right-hand sides, which is undoubtedly the most famous equilibrium condition in the allocation of consumption over time,[19] holds if, and only if, the wage rates are the same in the i^{th} and j^{th} periods.

Consider the implications of (43) for the optimal consumption path over time. I assume neutral time preference in the weak sense that all the U_i would be the same if all the C_i were the same. Then if equality held in (43), all the C_i would be the same if $r = 0$, and would tend to rise over time (ignoring differential wealth effects) if $r > 0$. Equality holds, however, only if the w_i were the same in all periods. But actual wage rates tend to rise with age until the mid-forties, fifties, or sixties, and then begin to decline. With that pattern for the w_i, (43) implies that if $r = 0$, the C_i would not be stationary, but would tend to decline with age until the peak w_i was reached, and then would tend to rise as the w_i fell (see Chart 2).[20]

The rate of fall and then rise of the C_i depends, of course, on the elasticities of substitution in consumption. In addition, the initial decline in C_i would be shorter and less steep and the subsequent rise would be longer, the larger r was (see Chart 2); for sufficiently large r, C_i might rise throughout.

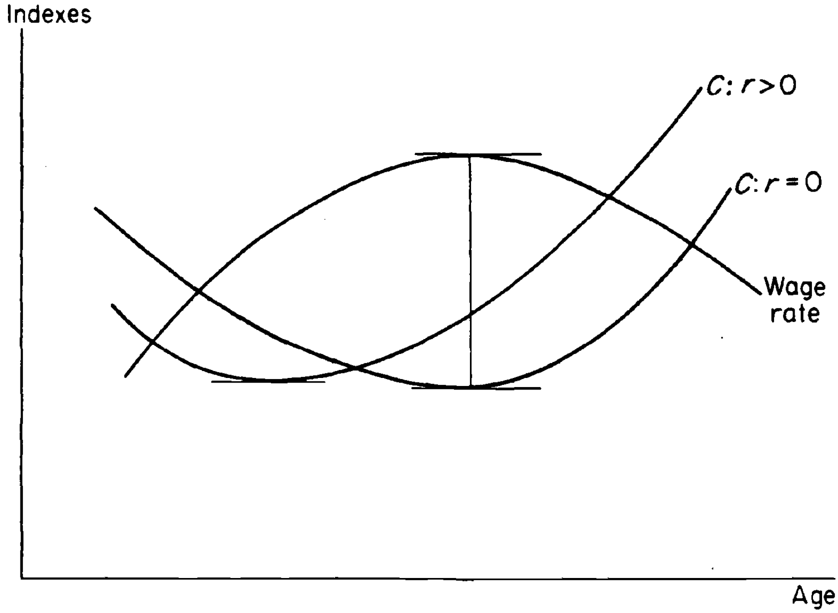
Since $\frac{t_c}{x}$, the ratio of consumption time to goods, would fall as the wage rate rose, and rise as it fell (see Chart 3), if C_i were constant, the absolute value of t_c would have the same pattern as this ratio. A fortiori, if $r = 0$, and if C_i declined as w_i rose and rose as w_i fell (see Chart 2), t_c would fall as w_i rose and rise as it fell (see Chart 3). If $r > 0$, C_i declines more briefly and less rapidly than when $r = 0$, and consequently, so would t_c; in particular, t_c would reach a minimum before w_i reached a maximum. Put differently, hours of work, t_w, would reach a maximum before the wage rate did. The difference between the peaks in t_w and w would be positively related to the size of r, and the elasticities of substitution between different C_i and C_j. Households faced with high interest rates, for example, should hit their peak hours of work earlier than otherwise similar households with low interest rates.

[19] Its derivation is presumably due to I. Fisher (see *The Theory of Interest*, New York, 1965, Chapters XII and XIII); it is also used in countless other studies: see, for example, J. Henderson and R. Quandt, *Microeconomics: A Mathematical Approach*, New York, 1971.

[20] I say "tend to" because of possible differential degrees of substitution between consumption in different periods. For example, high consumption in period l might so raise the marginal utility of consumption in period k as to cause the equilibrium value of C_k to exceed C_j, even though $w_j < w_k$. If the utility function is fully separable, this cannot occur.

CHART 2

Relations between Age, Wage Rates, and Commodity Consumption



It may appear that the Fisherian equality has simply been hidden and not replaced by the concentration on C instead of x. Indeed, equation (42) does imply a kind of Fisherian equality if the f terms are transposed to the left side to yield

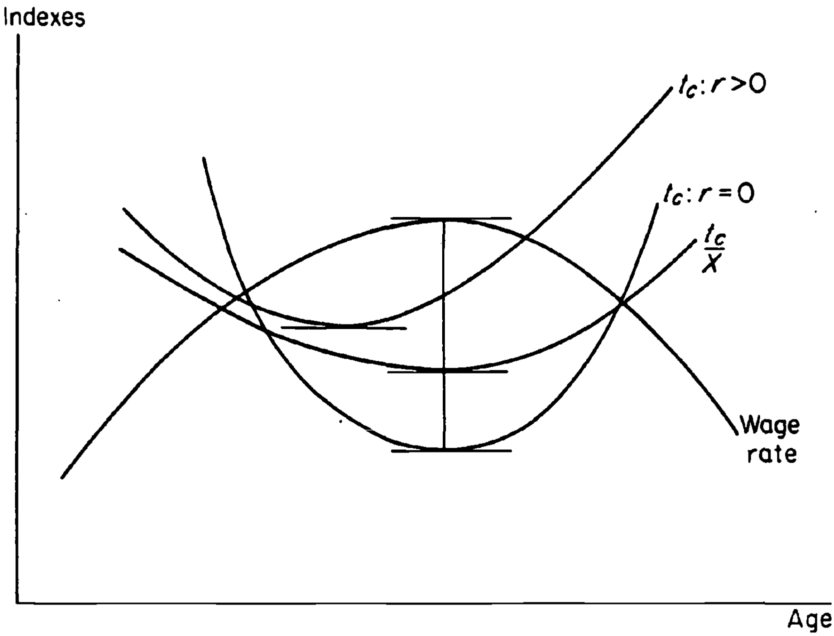
$$\frac{MU_{x_i}}{MU_{x_j}} = \frac{U_i f_{x_i}}{U_j f_{x_j}} = (1 + r)^{j-i}. \tag{44}$$

The term $U_i f_{x_i}$ is the marginal utility of an additional unit of x_i, and similarly for the j term. Equation (44) would seem to imply a horizontal path of the x_i if $r = 0$ and if time preference were neutral, the Fisherian result.

However plausible, this conclusion does not follow, and the Fisherian result cannot be saved. This is partly because the utility function depends directly on C and only indirectly on x, and partly because the path of x is also dependent on the production function f. If $r = 0$ and U implied neutral time preference with respect to the C, then the movement in C would tend to be inversely, and that in x/t_c directly,

CHART 3

Relations between Age, Wage Rates, and Time Spent in Consumption



related to the movement in w. The size of these respective movements depends on the elasticities of substitution between the C in U, and between x and t_c in f. The movement in C tends to make x inversely related to the movement in w, whereas that in x/t_c makes it directly related.

The actual movement in x, therefore, is determined by the relative strength of these opposing forces, that is, by the relative size of the elasticities of substitution in consumption and in production. The larger the latter elasticity, the more likely that x is directly related to w. Only if the elasticities were identical would the two substitutions offset each other, and would x be stationary with $r = 0$.[21] Of course, x (and C) are more likely to rise over time the higher r is.

Note that a rise in the consumption of goods with age, which is

[21] For further developments, see G. Ghez, *A Theory of Life Cycle Consumption*, Ph.D. dissertation, Columbia University, 1970, and G. Ghez and G. Becker, *The Allocation of Time and Goods over the Life Cycle*, New York, NBER, 1975.

frequently observed at least until age forty-five, can be explained without recourse to assumptions about time preference for the future, elastic responses to interest rate changes, or underestimation of future incomes. Neutral time preference, negligible interest rate responses, and perfect anticipation of the future could all be assumed if there were sufficiently easy substitution between time and goods in the production of commodities. The time path of goods consumption is not, however, a reliable guide to the path of true consumption (that is, of commodities) since the latter could well be inversely correlated with the former.

Investment in Human Capital

Instead of assuming that time can be allocated only between market labor force activity and nonmarket consumption activity, I now introduce a third category, investment in human capital. For the present an increased amount of human capital, measured by E, is assumed to affect only wage rates. Each person produces his own human capital by using some of his time and goods to attend "school," receive on-the-job training, etc. The rate of change in his capital equals the difference between his rate of production and the rate of depreciation on his stock.[22]

In symbols,

$$\phi_i = \psi_i(t_{e_i}, x_{e_i}), \tag{45}$$

where ϕ_i is the output of human capital in the i^{th} period, and t_{e_i} and x_{e_i} are the time and goods inputs, respectively. Then

$$E_{i+1} = E_i + \phi_i - dE_i, \tag{46}$$

where E_{i+1} is the stock at the beginning of the $i + 1$ period, and d is the rate of depreciation during a period. Each household maximizes the utility function in (32), subject to the production constraints in

[22] This model of human capital accumulation is very similar to and much influenced by that found in Y. Ben-Porath, "The Production of Human Capital and the Life Cycle of Earnings," *Journal of Political Economy,* August 1967, or in the addendum to this volume "Human Capital and the Personal Distribution of Income: An Analytical Approach," pp. 94–144.

(33), (45), and (46), and to the following time and goods "budget" constraints

$$t_{c_i} + t_{w_i} + t_{e_i} = t, \qquad i = 1, \ldots n \quad (47)$$

$$\sum_{i=1}^{n} \frac{x_i + x_{e_i}}{(1+r)^i} = \sum_{i=1}^{n} \frac{\alpha_i E_i t_{w_i} + v_i}{(1+r)^i}, \quad (48)$$

where $w_i = \alpha_i E_i$ and α_i is the payment per unit of human capital in period i. If, for simplicity, one assumes that ϕ_i depends only on t_{e_i} and that ψ_i is the same in all periods, and if the optimal solution has nonzero values of x_i, t_{c_i}, t_{w_i}, and t_{e_i}, the first order optimality conditions are

$$U_i f_{x_i} = \lambda \frac{1}{(1+r)^i} \qquad i = 1, \ldots n \quad (49)$$

$$U_i f_{t_i} = \lambda \frac{\alpha_i E_i}{(1+r)^i} \qquad i = 1, \ldots n \quad (50)$$

$$0 = \lambda \left[\frac{\alpha_i E_i}{(1+r)^i} - \sum_{j=i+1}^{n} \frac{\alpha_j t_{w_j}}{(1+r)^j} \frac{\partial E_j}{\partial t_{e_i}} \right]. \quad (51)$$

Equations (49) and (50) are essentially the same as (38) and (39). Therefore, investment in human capital, under the present assumptions, does not basically change the implications derived so far. For example, the time spent in consuming, t_c, would still tend to decline with age, reach a trough before the peak wage rate age, and then increase, and the time path of goods would still depend on the interest rate, and the elasticities of substitution in consumption and production. Two significant differences are, first, that the path of the wage rate is no longer given, but is determined by the path of the endogenous variable E_i. The wage rate would reach a peak before, at, or after the peak in α_i as E_i peaked before, at, or after α_i. Second, the behavior of t_w is no longer simply the complement of the behavior of t_c, since t_w also depends on t_e, which is determined by equation (51).

Equation (51) expresses the well-known equilibrium condition that the present value of the marginal cost of investing in human capital equals the present value of future returns. This equation clearly shows that the amount of time spent investing in human capital would tend to decline with age for two reasons. One is that the number of remaining periods, and thus the present value of future returns, would decline with age. The other is that the cost of investment would tend to rise with age as E_i rose because foregone earnings would rise.

Several interesting consequences follow from the tendency for t_{e_i} to fall as i increases. One is that hours of work, t_w, would be lower at younger ages and rise more rapidly than if there were no investment in human capital. Consequently, as long as t_{e_i} was positive, the peak in t_w would tend to come after the trough in t_c, and might even also come after the peak in w_i. However, since t_e declines with age, if it became sufficiently small by some age p before n, then for $i > p$, the behavior of t_w would be approximately the complement of the behavior of t_c.

If so much time at younger ages were put into investment in human capital that no time remained to allocate to work ($t_w = 0$), t_c and t_e would be complements at these ages. Marginal investment costs would not be measured by foregone earnings, but by the marginal value of time used in consumption, which would exceed foregone earnings (otherwise $t_w > 0$).

If $t_{w_i} = 0$, $i = 1, \dots q$, instead of equations (49) to (51), the first order optimality condition for $i = 1, \dots q$ would be

$$U_i f_{x_i} = \lambda \frac{1}{(1+r)^i} \qquad i = 1, \dots q \quad (52)$$

$$U_i f_{t_i} = s_i \qquad i = 1, \dots q \quad (53)$$

$$s_i = \lambda \sum_{j=i+1}^{n} \frac{\alpha_j t_{w_j}}{(1+r)^j} \frac{\partial E_j}{\partial t_{e_i}} \qquad i = 1, \dots q \quad (54)$$

where s_i is the marginal utility of an additional hour of time spent at consumption in the i^{th} period. If $U_i f_{t_i}$ is substituted for s_i in equation (54),

$$\frac{U_i f_{t_i}}{\lambda} = \sum_{j=i+1}^{n} \frac{\alpha_j t_{w_j}}{(1+r)^j} \frac{\partial E_j}{\partial t_{e_i}}, \qquad i = 1 \dots q \quad (55)$$

or the present value of the returns from an additional hour spent investing would equal not foregone earnings but the money equivalent of the marginal utility from an additional hour spent in consumption.

When equation (53) is divided by (52), and a substitution for s_i is made from (54), one has

$$\frac{f_{t_i}}{f_{x_i}} = \sum_{j=i+1}^{n} \frac{\alpha_j t_{w_j}}{(1+r)^{j-i}} \frac{\partial E_j}{\partial t_{e_i}}; \qquad i = 1 \dots q \quad (56)$$

the ratio of the marginal products of time and goods is not equated to the wage rate since no time is spent working, but to the monetary value of the marginal productivity of time used in investing. Even if w_i for $i < q$ were small, therefore, the production of commodities would be goods-intensive if the return to investment time were high.

Age and the Production Functions

By assuming that the production functions for commodities and human capital are the same at all ages, I have been able to analyze the different time and goods combinations at different ages in terms of differences in real wage rates and returns alone. Yet presumably as a person gains (or loses) experience, knowledge, and strength with age, the production possibilities available to him also change. This section analyzes the consequences of such changes for the optimal allocation of goods and time.

Let us concentrate on changes in the production functions for commodities, and assume that productive efficiency rises with age until a peak efficiency is reached, and then declines until age n. If the changes in efficiency were factor neutral, the production functions could be written as

$$ {}_jf = g_j f(t_{cj}, x_{cj}), \tag{57}$$

where the g_j are coefficients that rise at first and then decline. Equation (49) would become

$$U_i\, {}_if_{x_i} \equiv g_i U_i f_{x_i} = \lambda \frac{1}{(1+r)^i}, \tag{58}$$

equation (50) would become

$$U_i\, {}_if_{t_i} = g_i U_i f_{t_i} = \lambda \frac{\alpha_i E_i}{(1+r)^i}, \tag{59}$$

while equation (51) would be unaffected.

If equation (58) is divided by (59), the efficiency coefficients g_i drop out, and the optimal combination of time and goods depends, as before, only on the shape of f and $\alpha_i E_i$. Therefore, goods intensity rises until a peak is reached at the peak wage age, and then falls, and does not at all depend on the path of the s_i.

If $r = 0$ and w were rising with age, the decline in C with age would be greater than when production functions did not change if productive efficiency (measured by g) were falling with age because the marginal cost of producing C would rise faster with age; conversely if g were rising with age. The effect on the x and t_c is less definite and depends also on the elasticity of substitution in consumption because changes in the efficiency of producing C—the use of x and t_c per unit of C—can offset the change in C. If this elasticity exceeded unity, changes in efficiency would change x and t_c in the same direction as it changes C.

If changes in efficiency were not factor neutral but, say, changed the marginal product of consumption time more than that of goods ("goods-saving" change), there would be less incentive to substitute goods for time as wages rose if efficiency also rose. Therefore, production would not become as goods-intensive when wages and efficiency were rising, or as time-intensive when they were both falling. The converse would hold, of course, if changes in efficiency changed the marginal product of goods more than time.

Human Capital and Consumption

So far I have assumed that an increase in human capital directly only changes the productivity of time in the marketplace. Human capital might, however, also change the productivity of time and goods used in producing household consumption or in producing additional human capital itself.

Studies of investment in education and other human capital have been repeatedly criticized for ignoring the consumption aspects, although critics have been no more successful than others in treating these aspects in a meaningful way. One approach is to permit human capital to enter utility functions, but given the difficulties in measuring, quantifying, and comparing utilities, this does not seem too promising. An alternative is to assume that human capital "shifts" household production functions,[23] as in

$$C_i = {}_if(x_i, t_{c_i}; E_i). \tag{60}$$

The marginal effect of human capital on consumption in the i^{th} period can be defined as the marginal product or "shift" of C_i with respect to E_i:

$$MP_{e_i} = \frac{\partial C_i}{\partial E_i} = \frac{\partial {}_if}{\partial E_i} = {}_if_{e_i}. \tag{61}$$

[23] This approach is treated in considerable detail, both theoretically and empirically, by R. Michael, *The Effect of Education on Efficiency in Consumption*, New York, NBER, 1972, an outgrowth of a 1969 Ph.D. dissertation at Columbia.

The optimal allocation of time and goods can still be found by differentiating the utility function subject to the production functions and budget constraints. Equilibrium conditions (49) and (50) or (52) and (53) would be formally unaffected by the inclusion of E in the production of commodities. The equilibrium conditions with respect to t_e, the time spent investing in human capital, would, however, change from equation (51) to

$$\sum_{j=i+1}^{n} U_j f_{e_j} \frac{\partial E_j}{\partial t_{e_i}} = \lambda \left(\frac{\alpha_i E_i}{(1+r)^i} - \sum_{j=i+1}^{n} \frac{\alpha_j t_{w_j}}{(1+r)^j} \frac{\partial E_j}{\partial t_{e_i}} \right) \tag{62}$$

or

$$\frac{\alpha_i E_i}{(1+r)^i} = \sum_{j=i+1}^{n} \frac{U_j}{\lambda} f_{e_j} \frac{\partial E_j}{\partial t_{e_i}} + \sum_{j=i+1}^{n} \frac{\alpha_j t_{w_j}}{(1+r)^j} \frac{\partial E_j}{\partial t_{e_i}}. \tag{63}$$

A similar change would be produced in equation (54).

The term on the left-hand side of equation (63) is the present value of foregone earnings in period i—the cost of using more time in period i to produce human capital—and the terms on the right give the present value of the benefits. The second term on the right is the familiar present value of monetary returns, and gives the increase in wealth resulting from an additional investment in human capital in period i. The first term on the right is less familiar and measures the effect of additional investment on consumption. It essentially measures the present value of the reduction in goods and time required to produce a given basket of commodities resulting from increased investment in period i.[24]

[24] Since f is homogeneous of the first degree in x and t_c,

$$C_i \equiv f_{x_i} x_i + f_{t_i} t_{c_i}, \tag{1'}$$

then

$$\frac{\partial C_i}{\partial E_i} \equiv f_{e_i} \equiv \frac{\partial f_{x_i}}{\partial E_i} x_i + \frac{\partial f_{t_i}}{\partial E_i} t_{c_i}. \tag{2'}$$

Define

$$\tilde{f}_{x_i} \equiv \frac{\partial f_{x_i}}{\partial E_i} \Big/ f_{x_i} \quad \text{and} \quad \tilde{f}_{t_i} \equiv \frac{\partial f_{t_i}}{\partial E_i} \Big/ f_{t_i}. \tag{3'}$$

Then

$$f_{e_i} \equiv \tilde{f}_{x_i}(f_{x_i} x_i) + \tilde{f}_{t_i}(f_{t_i} t_{c_i}). \tag{4'}$$

Substituting from the equilibrium conditions (49) and (50) for f_{x_i} and f_{t_i} yields

$$f_{e_i} = \frac{\lambda}{U_i} \left(\tilde{f}_{x_i} \frac{x_i}{(1+r)^i} + \tilde{f}_{t_i} \frac{\alpha_i E_i t_{c_i}}{(1+r)^i} \right), \tag{5'}$$

Treated in this way, the effect of human capital on consumption becomes symmetrical to its effect on investment: the latter gives the monetary value of the stream of increased incomes, whereas the former gives the monetary value of the stream of reduced costs.

A few implications of the inclusion of the consumption effects of human capital can be noted briefly. Since they clearly raise the total benefits from investment, more time at each age would be spent investing than if these effects were nil. This in turn implies a greater likelihood of "corner" solutions, especially at younger ages, with the equilibrium conditions given by equations (52), (53), and an extension of (54)[25] being relevant. Moreover, there would now be justification for an assumption that efficiency in consumption and wage rates rise and fall together, because they would be the joint results of changes in the stock of human capital.

Some Extensions of the Analysis

It is neither realistic nor necessary to assume that wage rates are given, aside from the effects of human capital. The average wage rate and the number of hours a person works are generally related because of fatigue, differences between part-time and full-time opportunities,

---

and thus

$$\frac{U_i}{\lambda} f_{e_i} = \tilde{f}_{x_i} \frac{x_i}{(1+r)^i} + \tilde{f}_{t_i} \frac{\alpha_i E_i t_{c_i}}{(1+r)^i}. \qquad (6')$$

The terms $\tilde{f}_{x_i}$ and $\tilde{f}_{t_i}$ equal the percentage reductions in goods and time respectively in period i required to produce a given C_i resulting from the "shift" in f caused by a unit increase in E_i. Hence, the full term on the right-hand side of (6′) gives the present value of the savings in goods and time in period i required to achieve a given amount of C_i. Consequently,

$$\sum_{j=i+1}^{n} \frac{U_j}{\lambda} f_{e_j} \frac{\partial E_j}{\partial t_{e_i}} = \sum_{j=i+1}^{n} \frac{\tilde{f}_{x_j} x_j}{(1+r)^j} + \frac{\tilde{f}_{t_j} \alpha_j E_j t_{c_j}}{(1+r)^j} \frac{\partial E_j}{\partial t_{e_i}} \qquad (7')$$

gives the full present value of the savings in goods and time resulting from additional investment in human capital in period i.

[25] With consumption effects, equation (54) is replaced by

$$\sum_{j=i+1}^{n} U_j \frac{\partial E_j}{\partial t_{e_i}} - s_i + \lambda \sum_{j=i+1}^{n} \frac{\alpha_j t_{w_j}}{(1+r)^j} \frac{\partial E_j}{\partial t_{e_i}} = 0,$$

or

$$\frac{s_i}{\lambda} = \sum_{j=i+1}^{n} \frac{U_j}{\lambda} f_{e_j} \frac{\partial E_j}{\partial t_{e_i}} + \sum_{j=i+1}^{n} \frac{\alpha_j t_{w_j}}{(1+r)^j} \frac{\partial E_j}{\partial t_{e_i}}. \qquad (54')$$

fixed costs of working, overtime provisions, and so forth. Our analysis can easily incorporate an effect of t_w on w, as in

$$w_i = w_i(t_{w_i}), \tag{64}$$

or even more generally in

$$w_i = w_i(t_{w_i}, t_{w_{i-1}}, \dots t_{w_1}) \tag{65}$$

if on-the-job learning is to be analyzed separately from other human capital. Marginal, not average, wage rates are the relevant measures of the cost of time and would enter the equilibrium conditions.[26]

It would also be more realistic to consider several commodities at any moment in time, each having its own production function and goods and time inputs. This could easily be done by introducing the utility function

$$U = U(C_{1l}, \dots C_{1n}, C_{2l}, \dots C_{2n}, \dots C_{ml}, \dots C_{mn}), \tag{66}$$

where C_j is the amount of the j^{th} commodity consumed in the i^{th} period. This function would be maximized subject to separate production functions for each commodity (and perhaps in each period) and to the budget constraints. One of the main implications is that when wage rates are relatively high, not only is the production of each commodity relatively goods-intensive, but consumption shifts toward relatively goods-intensive commodities and away from time-intensive commodities. The latter (such as children or grandchildren) would be consumed more at younger and older ages if wage rates or more generally the cost of time rose at younger ages and fell eventually; conversely, goods-intensive commodities would be consumed more at middle ages. These age patterns in the consumption of time and goods-intensive commodities strengthen the tendency for consumption time to fall initially and for goods to rise initially with age.

The accumulation of human capital might also "shift" the production function used to produce human capital itself since investors with much human capital might well be more productive than those with little. This has been discussed elsewhere,[27] and I only mention here

[26] For example, if equation (64) is the wage rate function, equation (65) would be replaced by

$$U_i f_{t_i} = \lambda \left(\frac{\alpha_i E_i}{(1+r)^i} + \frac{\partial \alpha_i}{\partial t_{w_i}} \frac{t_{w_i} E_i}{(1+r)^i} \right). \tag{65'}$$

[27] See Ben-Porath, *op. cit.*, and addendum to this volume "Human Capital and the Personal Distribution of Income: An Analytical Approach," pp. 94–144.

one implication. The tendency for the amount invested to decline with age would be somewhat retarded because investment would be encouraged as capital was accumulated, since time would become more productive and this would offset the effect of its becoming more costly.

The allocation over a lifetime should be put in a family context, with the decisions of husbands, wives, and possibly also children interacting with each other. For example, if wives' wage rates are more stationary than their husbands', the analysis in this paper predicts that the labor force participation of married women would be relatively high at younger and older ages, and relatively low at middle ages, precisely what is observed. A similar result would follow if the productivity in consumption of married women's time is higher at middle ages because child rearing is time-intensive. The analysis developed here seems capable of throwing considerable light on the differential labor force participation patterns by age of husbands and wives.[28]

Empirical Analysis

Some implications of this model have been tested by the author with data from the 1960 Census 1/1000 sample giving earnings, hours, and weeks worked, cross-classified by age, sex, race, and education.[29]

2. The Incentive to Invest

Number of Periods

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 and the next one, 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, in-

[28] This has been confirmed in several studies since this was written; see A. Leibowitz, "Women's Allocation of Time to Market and Non-Market Activities," Ph.D. dissertation, Columbia University, 1972; or H. Ofek, "Allocation of Goods and Time in a Family Context," Ph.D. dissertation, Columbia University, 1971; or J. Smith, "A Life Cycle Family Model, NBER Working Paper 5, 1973.

[29] The results are published in Ghez and Becker, *op. cit.*, Chapter 3.

cluding much that has been either ignored or given ad hoc interpretations. The discussion covers many topics, starting with the life span of activities and ending with a theory of the distribution of earnings.

An increase in the life span of an activity would, other things being equal, increase the rate of return on the investment made in any period. The influence of life span on the rate of return and thus on the incentive to invest is important and takes many forms, a few of which will now be discussed.

The number of periods is clearly affected by mortality and morbidity rates; the lower they are, the longer is the expected life span and the larger is 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 probably increased the rates of return on investment in human capital,[30] thereby encouraging such investment.[31] This conclusion is independent 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 the investment in health.

A relatively large fraction of younger persons are in school or on-the-job training, change jobs and locations, and add to their knowledge of economic, political, and social opportunities. The main explanation may not be that the young are relatively more interested in learning, 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 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. Indeed, there would be a greater incentive even if age had no effect on capabilities, responsibilities, and attitudes.

The ability to collect returns over more years would give young

30 I say *probably* because rates of return are adversely affected (via the effects on marginal productivity) by the increase in labor force that would result from a decline in death and sickness. If the adverse effect were sufficiently great, their decline would reduce rates of return on human capital. I am indebted to my wife for emphasizing this point.

31 The relation between investment in training and length of life is apparently even found in the training of animals, as evidenced by this statement from a book I read to my children: "Working elephants go through a long period of schooling. Training requires about ten years and costs nearly five thousand dollars. In view of the animal's long life of usefulness [they usually live more than sixty years], this is not considered too great an investment" (M. H. Wilson, *Animals of the World,* New York, 1960).

persons a much greater incentive to invest even if the internal rate of return did not decline much with age. The internal rate can be seriously misleading here, as the following example indicates. If $100 invested at any age yielded $10 a year additional income forever, the rate of return would be 10 per cent at every age, and there would be no special incentive to invest at younger ages if only the rate of return were taken into account. Consider, however, a cohort of persons aged eighteen deciding when to invest. If the rate of return elsewhere were 5 per cent and if they invested immediately, the present value of the gain would be $100. If they waited five years, the present value of the gain, i.e., as of age eighteen, would only be about $78, or 22 per cent less; if they waited ten years, the present value of the gain would be under $50, or less than half. Accordingly, a considerable incentive would exist for everyone to invest immediately rather than waiting. In less extreme examples some persons might wait until older ages, but the number investing would tend to decline rapidly with age even if the rate of return did not.[32]

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.[33] Suppose that investment in human capital raised earnings for p periods only, where p varied between 0 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 were correct, the difference between the amount

32 One clear application of these considerations can be found in studies of migration, where some writers have rejected the importance of the period of returns because migration rates decline strongly with age, at least initially, while rates of return (or some equivalent) decline slowly (see the otherwise fine paper by L. Sjaastad, "The Costs and Returns of Human Migration," *Investment in Human Beings*, pp. 89–90). My analysis suggests, however, that persons with a clear gain from migration have a strong incentive to migrate early and not wait even a few years. Since the persons remaining presumably have either no incentive or little incentive to migrate, it is not surprising that their migration rates should be much lower than that of all persons.

33 This test was suggested by George Stigler's discussion of the effect of different autocorrelation patterns on the incentive to invest in information (see "The Economics of Information," *Journal of Political Economy*, June 1961, and "Information in the Labor Market," *Investment in Human Beings*, pp. 94–105).

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;[34] 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 find investments that increase productivity in several activities. A woman wants her investment to be useful both in her roles 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," which is so often stressed in discussions of the division of labor,[35] but also because a larger market would *induce* a greater investment in skills.[36] Put differently, the usual analysis of the division of labor stresses that efficiency, and thus wage rates, would be

[34] This example is from Stigler, "The Economics of Information," *Journal of Political Economy*, June 1961.

[35] See, for example, A. Marshall, *Principles of Economics*, New York, 1949, Book IV, Chapter ix.

[36] 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 above distinction between the effect of an increase in the market on practice and on the incentive to invest would then simply be that the incentive to invest in human capital is increased even aside from the effect of practice on earnings.

greater the larger the market, and ignores the potential earnings period in any activity, while mine 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 latter, that is, to the influence of market size on the incentive to invest in skills.

Wage Differentials and Secular Changes

According to equation (30), the internal rate of return depends on the ratio of the return per unit of 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,[37] although no one has adequately discussed their relative merits. Since marginal productivity analysis relates the derived demand for any class of workers to the ratio of their wages to those of other inputs,[38] 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. 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

[37] See A. M. Ross and W. Goldner, "Forces Affecting the Interindustry Wage Structure," *Quarterly Journal of Economics*, May 1950; P. H. Bell, "Cyclical Variations and Trend in Occupational Wage Differentials in American Industry since 1914," *Review of Economics and Statistics*, November 1951; F. Meyers and R. L. Bowlby, "The Interindustry Wage Structure and Productivity," *Industrial and Labor Relations Review*, October 1953; G. Stigler and D. Blank, *The Demand and Supply of Scientific Personnel*, New York, NBER, 1957, Table 11; P. Keat, "Long-Run Changes in Occupational Wage Structure, 1900–1956," *Journal of Political Economy*, December 1960.

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

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,[39] 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 the analysis here 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 differences and costs. Clearly a secular decline in wage ratios would not be 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.[40]

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 reexamination of the evi-

[39] 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.

[40] Keat's data for 1906 and 1953 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 (*ibid.*); I computed the change in the real standard deviation from data made available to me by Keat.

dence,[41] 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 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.[42]

Risk and Liquidity

An informed, rational person would invest only if the expected rate of return were 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 has always 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 knowledge re-

41 For those believing that the qualitative evidence overwhelmingly indicates a continuous 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,* Modern Library edition, New York, 1937, p. 106). Since economists tend to believe that law and most other liberal professions are now overcompensated relative to nonprofessional work "in point of pecuniary gain," the return to professional work could not have declined continuously if Smith's observations were accurate.

42 Some quantitative evidence for the United States is discussed in Chapter VI, section 2.

quired is 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.[43] 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,[44] and young persons are supposed to be especially prone to overestimate their ability and chance of good fortune.[45] Were this view correct, a human investment that 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 above). Indeed, an alternative explanation of reactions to large gains has already appeared.[46]

Capital Markets and Knowledge

If investment decisions responded 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 on nonhuman capital, however, because of financing diffi-

43 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" and "the circumstances by which the earnings are determined are less capable of being foreseen [than those for machinery]" (*Principles of Economics,* p. 571). In section 4 of Chapter IV some quantitative estimates of the uncertainty in the return to education are presented.

44 Note that our argument above implied that investors in human capital would be younger.

45 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" (*Wealth of Nations,* 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" (*Principles of Economics,* p. 554).

46 See M. Friedman and L. J. Savage, "The Utility Analysis of Choices Involving Risks," reprinted in *Readings in Price Theory,* G. J. Stigler and K. Boulding, eds., Chicago, 1952.

culties 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 that 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.[47]

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 \$1000, and if all costs were foregone earnings, investors would have \$200 of earnings to spend; if all were direct costs, they would initially have \$1000 to spend, but just \$200 would remain after

[47] 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 the loans.

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 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 also seem 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 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 nonhuman 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.

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.

3. Some Effects of Human Capital

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). The previous discussion 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 (Chapter II, section 1) that on-the-job training would steepen age-earnings profiles, and the analysis of section 1 of this chapter generalizes the argument to all human capital. For 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.[48] Likewise, investment in human capital would make the profile more concave.[49]

[48] According to equation (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.

[49] Following the notation of the previous footnote, Y would be more concave than X if and only if

$$\Delta\left(\frac{\Delta Y_j}{\Delta j}\right) - \Delta\left(\frac{\Delta X_j}{\Delta j}\right) = \Delta\left(\frac{r_j C_j}{\Delta j}\right) - \Delta\left(\frac{\Delta C_j}{\Delta j}\right) < 0.$$

The first term on the right is certain to be negative, at least eventually, because

2. In recent years students of international trade theory have been somewhat shaken by findings that the United States, said to have a relative scarcity of labor and an 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.[50]

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.[51]

3. Several studies have tried to estimate empirically the elasticity of substitution between capital and labor. Usually a ratio of the input of physical capital (or output) 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.[52] 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

---

both r_j and C_j would eventually decline, while the second term would be positive because C_j would eventually decline at a decreasing rate. Consequently, the inequality would tend to hold and the earnings profile in Y would be more concave than that in X.

50 See I. Kravis, "Wages and Foreign Trade," *Review of Economics and Statistics,* February 1956.

51 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,* November 1956.

52 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,* June 1961, pp. 261–270; 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.

could give a seriously biased picture of the effect on wages of factor proportions.[53]

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.[54] Instead of only benefiting from activities by others, the average earner is made a prime mover of development through the investment in himself.[55]

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 between 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 consists of particular kinds 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

[53] Minasian's argument (in his article cited above, 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 stated that he intended to make this condition explicit.) If, on the other hand, human and physical capital were perfect substitutes, I have shown (in an unpublished memorandum) that 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 generally to determine the direction of bias in this case.

[54] The major figure here is T. W. Schultz. Of his many articles, see especially "Education and Economic Growth," in *Social Forces Influencing American Education,* Sixtieth Yearbook of the National Society for the Study of Education, Chicago, 1961, Part II, Chapter 3.

[55] 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 would be less than if the cost of human capital, like that of physical capital, was not deducted from national income.

results, or by earnings themselves.[56] 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 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 relevant and should not be ignored.

A compromise might be reached through defining ability by earnings only when several variables have 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, luck, 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, it has been shown that, after the investment period, earnings (Y) can be simply approximated by

$$Y = X + rC, \tag{67}$$

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.[57]

In most capital markets the amount invested is not the same for everyone nor 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.[58] Since marginal

[56] Let me state again that the word "earnings" stands for real earnings, or the sum of monetary earnings and the monetary equivalent of psychic earnings.

[57] 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.

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

and average rates are presumably positively correlated[59] 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,[60] a correlation with several important implications.

One is that the tendency for abler persons to migrate, continue their education,[61] and generally invest more in themselves can be explained without recourse to an assumption that noneconomic 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.[62]

The main implication, however, is in 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.[63] Pigou's main suggestion—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 relation to the main-

[59] 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.

[60] 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" (*Principles of Economics*, p. 312).

[61] The first is frequently alleged (see, for example, *ibid.*, p. 199). Evidence on the second is discussed in Chapter IV, section 2.

[62] See my "An Economic Analysis of Fertility," in *Demographic and Economic Change in Developed Countries*, Special Conference 11, Princeton for NBER, 1960.

[63] See A. C. Pigou, *The Economics of Welfare*, 4th ed., London, 1950, Part IV, Chapter ii.

stream of economic thought.[64] 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 (67) 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.[65] 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 were statistically independent of ability.[66]

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

[65] J. Mincer ("Investment in Human Capital and Personal Income Distribution," *Journal of Political Economy,* 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 Mincer is followed in assuming that everyone was equally able, that schooling was the only investment, and that the cost of the n^{th} 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

$$\begin{aligned}
C_1 &= Y_0 \\
C_2 &= Y_1 = Y_0(1 + r) \\
C_3 &= Y_2 = Y_1(1 + r) = Y_0(1 + r)^2 \\
C_n &= Y_{n-1} = \cdots = Y_0(1 + r)^{n-1},
\end{aligned}$$

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 equation (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.

[66] 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,* March 1936, p. 3).

It has been 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.[67] 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 from investment in human capital—the smaller rC relative to X—the more the distribution of earnings would be dominated by the distribution of X. Higher-skill categories have a greater average investment in human capital and thus presumably a larger 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.[68]

Equation (67) holds only when investment costs are small, which tends to be true at later ages, say, after age thirty-five. Net earnings at earlier ages would be given by

$$Y_j = X_j + \sum_{0}^{j-1} r_i C_i + (-C_j), \tag{68}$$

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_j. The distribution of $-C_j$ would be an important determinant

[67] Craig (*ibid.*, 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.

[68] 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 14, 25, or 6). Indeed, the developmental approach to child rearing argues that earnings would be close to zero if there were 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 the distribution of Y_j since investment is large at these ages. Hence the analysis would predict a smaller (positive) skewness at younger than at older ages partly because X would be more important relative to $\Sigma r_i C_i$ at younger ages and partly because the presumed negative correlation between $-C_j$ and $\sum_{0}^{j-1} r_i C_i$ would counteract the positive correlation between r_i and C_i.

A simple analysis of the incentive to invest in human capital seems capable of explaining, therefore, not only why the overall 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.

Addendum: Education and the Distribution of Earnings: A Statistical Formulation[69]

A Statistical Formulation

The contribution of human capital to the distribution of earnings could be easily calculated empirically if the rates of return and investments in equation (1) were known.[70] Although information on investment in human capital has grown significantly during the last few years, it is still limited to aggregate relations for a small number of countries. Much more is known about one component of these investments; namely, the period of time spent investing, as given, for example, by years of schooling.

To utilize this information we have reformulated the analysis to bring out explicitly the relation between earnings and the investment period. The principal device used is to write the cost of the j^{th} "year" of investment to the i^{th} person as the fraction k_{ij} of the earnings that would be received if no investment was made during that year. If for

[69] Reprinted from pp. 363–369 of an article by G. S. Becker and B. R. Chiswick in *American Economic Review*, May 1966.

[70] Equation (1):

$$E_i = X_i + \sum_{j=1}^{m} r_{ij} C_{ij},$$

where C_{ij} is the amount spent by the i^{th} person on the j^{th} investment, r_{ij} is his rate of return on this investment, and X_i is the effects of the original capital.

convenience r_{ij} in equation (1) is replaced by $\bar{r}_j + r_{ij}^*$, where $\bar{r}_j$ is the average rate of return on the j^{th} investment and r_{ij}^* is the (positive or negative) premium to the i^{th} person resulting from his (superior or inferior) personal characteristics, then it can be shown that equation (1) could be rewritten as

$$E_i = X_i[1 + k_{i1}(\bar{r}_1 + r_{i1}^*)][1 + k_{i2}(\bar{r}_2 + r_{i2}^*)] \cdots [1 + k_{in_i}(\bar{r}_{n_i} + r_{in_i}^*)] \quad (69)$$

where n_i is the total investment period of the i^{th} person.[71] If the effect of luck and other such factors on earnings is now incorporated within a multiplicative term e^{u_i}, the log transform of equation (69) is

$$\log E_i = \log X_i + \sum_{j=1}^{n_i} \log [1 + k_{ij}(\bar{r}_j + r_{ij}^*)] + u_i. \quad (70)$$

By defining $X_i = \overline{X}(1 + \alpha_i)$, where α_i measures the "unskilled" personal characteristics of the i^{th} person, and $k_{ij} = \bar{k}_j + t_{ij}$, where $\bar{k}_j$ is the average fraction for the j^{th} investment, and by using the relation

$$\log [1 + k_{ij}(\bar{r}_j + r_{ij}^*)] \cong k_{ij}(\bar{r}_j + r_{ij}^*), \quad (71)$$

equation (70) could be written as

$$\log E_i \cong a + \sum_{j=1}^{n_i} \bar{r}_j' + v_i, \quad (72)$$

where $a = \log \overline{X}$, $\bar{r}'_j = \bar{k}_j\bar{r}_j$, and

$$v_i = \log (1 + \alpha_i) + \sum_j k_{ij}r_{ij}^* + \sum_j t_{ij}\bar{r}_j + u_i. \quad (73)$$

The term v_i largely shows the combined effect on earnings of luck and ability. If the $\bar{r}'_j$ was the same for each period of investment, the equation for earnings is simply

$$\log E_i \cong a + \bar{r}'n_i + v_i. \quad (74)$$

If $\bar{r}'$, the average rate of return adjusted for the average fraction of earnings foregone, and the investment period n_i were known, equation

71 The interested reader can find a proof for a somewhat special case in footnote 65 above.

(74) could be used to compute their contribution to the distribution of earnings. For example, they would jointly "explain" the fraction

$$R^2 = (\bar{r}')^2 \frac{\sigma^2(n)}{\sigma^2(\log E)} \qquad (75)$$

of the total inequality in earnings, where $\sigma^2(n)$ is the variance of investment periods, and $\sigma^2(\log E)$ is the variance of the log of earnings, the measure of inequality in earnings.[72] Ability and luck together would "explain" the fraction $\sigma^2(v)/\sigma^2(\log E)$, and the (perhaps negative) remainder of the inequality in earnings would be "explained" by the covariance between ability, luck, and the investment period.

Even equations (72) and (74), simplified versions of (69), make excessive demands on the available data. For one thing, although the period of formal schooling is now known with tolerable accuracy for the populations of many countries, only bits and pieces are known about the periods of formal and informal on-the-job training, and still less about other kinds of human capital. Unfortunately, the only recourse at present is to simplify further: by separating formal schooling from other human capital, equation (72) becomes

$$\log E_i = a + \sum_{j=1}^{q_i} \bar{r}'_j S_j + v'_i, \qquad (76)$$

where $\bar{r}'_1$ is the adjusted average rate of return on each of the first S_1 years of formal schooling, $\bar{r}'_2$ is a similar rate on each of the succeeding S_2 years of formal schooling, etc.;

$S_i = \sum_1^{q_i} S_j$ is then the total formal schooling years of the i^{th} person, and

$$v'_i = v_i + \sum \bar{r}'_k T_k \qquad (77)$$

includes the effect of other human capital.

A second difficulty is that although an almost bewildering array of rates of return have been estimated in recent years, our empirical analysis requires many more. Additional estimates could be developed by making equation (76) do double duty: first it could be used to estimate the adjusted rates and only then to show the contribution of

[72] Note that this measure, one of the most commonly used measures of income inequality, is not just arbitrarily introduced but is derived from the theory itself.

schooling, including these rates, to the distribution of earnings. If the S_j and v' were uncorrelated, an ordinary least squares regression of log E on the S_j would give unbiased estimates on these rates, and, therefore, of the contribution of schooling. If, however, the S_j and v' were positively or negatively correlated, the estimated rates would be biased upward or downward, and so would the estimated direct contribution of schooling.

Some components of v' are probably positively and others are negatively correlated with years of schooling, and the net bias, therefore, is not clear a priori. It is not unreasonable to assume that α_i and u_i in equation (73) are only slightly correlated with the S_j. The r_{ij}^* term in (73), on the other hand, would be positively correlated with the S_j[73] since the theory developed earlier suggests that persons of superior ability and other personal characteristics would invest more in themselves. Some empirical evidence indicates a positive correlation between years of schooling and the amount invested in other human capital.[74] The term v' depends, however, on the correlation between years of schooling and years invested in other human capital, a correlation which might well be negative. Certainly persons leaving school early begin their on-the-job learning early, and possibly continue for a relatively long time period (see fn. 75). Finally, one should note that random errors in measuring the period of schooling would produce a negative correlation between the measured period and v'. Although the correlations between the S_j and these components of v' go in both directions and thus to some extent must offset each other, a sizable error probably remains in estimating the adjusted rates and the contribution of schooling to the distribution of earnings.

Empirical Analysis

The sharpest regional difference in the United States in opportunities and other characteristics is between the South and non-South, and Table 1 presents some results of regressing the log of earnings on years of schooling separately in each region for white males at least age twenty-five. Adjusted average rates of return have been estimated by these regressions separately for low, medium, and high education levels. As columns 1, 2, 6, 7, and 8 indicate, the adjusted rates at each of these school levels and the variances in the log of earnings and in years of

[73] That is, unless a negative correlation between k_{ij} and r_{ij}^* was sufficiently strong.

[74] See p. 167.

TABLE 1

Results of Regressing Natural Log of Earnings on Education for 1959 Earnings of White Males Aged 25 to 64 in the South and Non-South

	Variance of Natural Log of Earnings (1)	Variance of Education (2)	Average Natural Log of Earnings[c] (3)	Average Education (4)	Inter-cepts[c] (5)	Adjusted Rates of Return[a] (standard errors)[b]: Low Education (6)	Medium Education (7)	High Education (8)	Adjusted Coefficient of Determination (R^2) / Unadjusted R^2 (9)	Residual Variance in Natural Log of Earnings (10)
Non-South	.42	11.28	1.66	10.78	1.09	.05	.06	.08	.07	.39
					(.67)	(.09)	(.06)	(.06)	.11	
South	.55	15.23	1.43	9.96	.66	.07	.09	.09	.16	.46
					(.50)	(.08)	(.07)	(.06)	.20	

Source: *United States Census of Population:* 1960, *Subject Reports-Occupation by Earnings and Education*, Bureau of the Census, Washington, D.C., 1963, Tables 2 and 3.

[a] "Low" education is defined as 0–8 years of school completed, "medium" as 8–12 years, and "high" as more than 12 years.

[b] In calculating the standard errors and the adjusted coefficients of determination, the number of degrees of freedom was assumed to equal the number of cells minus the number of parameters estimated. The true number is clearly somewhat greater than this.

[c] Earnings measured in thousands of dollars.

schooling are all a fair amount larger in the South. Moreover, these differences in schooling and rates exceed the difference in earnings, for as column 9 shows, the coefficient of determination, or the fraction of the variance in the log of earnings "explained" by schooling, is considerably higher in the South. The regional difference in earnings does not entirely result from schooling, however, for column 10 shows that the "residual" inequality in earnings is also larger in the South.

These results can be given an interesting interpretation within the framework of the theory presented in the addendum "Human Capital and the Personal Distribution of Income: An Analytical Approach." The greater inequality in the distribution of schooling in the South is presumably a consequence of the less equal opportunity even for whites there and would only be strengthened by considering the differences in schooling between whites and nonwhites. The higher adjusted rates of return in the South[75] are probably related to the lower education levels there, shown in column 4, which in turn might be the result of fewer educational opportunities.

The residual v' is heavily influenced by the distributions of luck and ability, which usually do not vary much between large regions. Therefore, greater rates of return and inequality in the distribution of schooling would go hand in hand not only with a greater absolute but also with a greater relative contribution of schooling to the inequality in earnings. The residual is also influenced, however, by investment in other human capital. Since the rates of return to and distribution of these investments would be influenced by the same forces influencing schooling, the absolute, but not relative, contribution of the residual to the inequality in earnings would tend to be greater when the absolute contribution of schooling was greater. Consequently, our theory can explain why both the coefficient of determination and the residual variance in earnings are greater in the South.

In order to determine whether these relations hold not only for the most extreme regional difference in the United States but also for

[75] Higher rates of return to whites in the South have been found when estimated by the "present value" method from data giving earnings classified by age, education, and other variables (see G. Hanoch, "An Economic Analysis of Earnings and Schooling," *Journal of Human Resources*, 2, Summer 1967, pp. 310–329). Although estimates based on the present value method are also biased upward by a positive correlation between ability and schooling and downward by errors in measuring school years, they are less sensitive to the omission of other human capital (see this volume, pp. 167–168). Consequently, the fact that Hanoch's estimates are almost uniformly higher than ours (after adjustment for the k_i) suggests, if anything, a negative correlation between school years and the years invested in other human capital. This could also explain why Hanoch's rates tend to decline with increases in years of schooling while ours tend to rise.

more moderate differences, similar regressions were calculated for all fifty states. To avoid going into details at this time let us simply report that the results strongly confirm those found at the extremes: there is a very sizable positive correlation across states between inequality in adult male incomes, adjusted rates of return, inequality in schooling, coefficients of determination, and residual inequality in incomes, while they are all negatively related to the average level of schooling and income. Whereas only about 18 per cent of the inequality in income within a state is explained, on the average, by schooling, the remaining 82 per cent explained by the residual, about one-third of the differences in inequality between states is directly explained by schooling, one-third directly by the residual, and the remaining one-third by both together through the positive correlation between them.

Similar calculations have also been made for several countries having readily available data: United States, Canada, Mexico, Israel, and Puerto Rico (treated as a country). Again there is a strong tendency for areas with greater income inequality to have higher rates of return, greater schooling inequality, higher coefficients of determination, and greater residual inequality. While there is also a tendency for poorer countries to have lower average years of schooling, greater inequality in income, etc., there are a couple of notable exceptions. For example, Israel, for reasons rather clearly related to the immigration of educated Europeans during the 1920s and 1930s, had unusually high schooling levels and low inequality in earnings until the immigration of uneducated Africans and Asians after 1948 began to lower average education levels and raise the inequality in earnings.

Addendum: Human Capital and the Personal Distribution of Income: An Analytical Approach*

1. Introduction

Interest among economists in the distribution of income has as long a history as modern economics itself. Smith, Mill, Ricardo, and others recognized that many problems of considerable economic importance partly turned on various aspects of income distribution. Although they defined poverty, for example, in absolute terms, they also recognized that each generation's "poor" are mainly those significantly below the average income level. In addition to poverty, the degree of opportu-

* Originally published as Woytinsky Lecture, University of Michigan, 1967.

nity, aggregate savings, and investment, the distribution of family sizes and the concentration of private economic power were believed to be related to income distribution.

How does one explain then that in spite of the rapid accumulation of empirical information and the persisting and even increasing interest in some of these questions, such as poverty, economists have somewhat neglected the study of personal income distribution during the past generation? In my judgment the fundamental reason is the absence, despite ingenious and valiant efforts, of a theory that both articulates with general economic theory and is useful in explaining actual differences among regions, countries, and time periods. By emphasizing investment in human capital one can develop a theory of income distribution that satisfies both desiderata. This essay focuses on the relation between investment in human capital and the distribution of earnings and other income. The discussion is theoretical and no systematic effort is made to test the theory empirically. I expect to report on some quantitative tests in a future publication.

The next section sets out the basic theory determining the amount invested in human capital by a "representative" person, and shows the relation between earnings, investments, and rates of return. Essentially all that is involved is the application to human capital of a framework traditionally used to analyze investment in other capital, although several modifications are introduced. Section 3 of the essay shifts the attention from a single person to differences among persons, and shows how the distribution of earnings and investments are determined by the distributions of ability, tastes, subsidies, wealth, and other variables.

Section 4 uses the framework developed in sections 2 and 3 to analyze the effects on the distribution of earnings of an increase in the equality of opportunity, of a more efficient market for human capital, of the use of tests and other "objective" criteria to ration investments in human capital, and of legislation requiring a minimum investment in human capital. Section 5 extends the discussion to the distribution of property income, and suggests why such income, both inherited and self-accumulated, is more unequally distributed than earnings. Section 6 summarizes the discussion and adds a few conclusions, and 7 is a mathematical appendix.

2. *Optimal Investment in Human Capital*

a. The model. I have shown elsewhere that what I call the "net" earnings of a person at any age t (E_t) approximately equals the earnings he would have at t if no human capital had been invested in him (X_t) plus the total returns to him at t on investments made in him earlier (k_t) minus the cost to him of investments at t (C_t), as in

$$E_t = X_t + k_t - C_t. \tag{78}$$

Total returns depend on the amounts invested and their rates of return; for example, if returns on each investment were the same at all ages during the labor force period,[76] total returns would be the sum of the products of the amounts invested and their rates of return, adjusted for the finiteness of the labor force period. Equation (78) could then be written as

$$E_t = X_t + \sum_{j=1}^{n} r_{t-j} f_{t-j} C_{t-j} - C_t, \tag{79}$$

where r_{t-j} is the rate of return on capital invested at $t - j$ and f_{t-j} is the finite life adjustment. I applied this analysis to various problems, including the shapes of age-earnings and age-wealth profiles, the relation between unemployment and on-the-job training, the so-called Leontief paradox, and several others.[77]

I suggested that differences in the total amounts invested by different persons are related to differences in the rates of return obtainable, a suggestion that can explain why white urban males with high IQs acquire more education than others, or why the division of labor is limited by the extent of the market.[78] I did not, however, systematically develop a framework to explain why rates of return and investments differ so greatly among persons. This essay tries to develop such a framework. This not only provides a rigorous justification for these suggestions in *Human Capital,* but also begins to provide an explanation of the personal distribution of earnings.

The term X_t in equations (78) and (79) represents the earnings of a person that are unrelated to human capital invested in him, and are presumably, therefore, largely independent of his current choices. Particularly in developed economies but perhaps in most, there is sufficient investment in education, training, informal learning, health, and just plain child rearing that the earnings unrelated to investment in human capital are a small part of the total. Indeed, in the develop-

76 This is the "one-hoss shay" assumption applied to human capital.

77 See this volume, Chapters II–III.

78 See pp. 74–75, 157–166, 169–181.

mental approaches to child rearing, all the earnings of a person are ultimately attributed to different kinds of investments made in him.[79] Consequently, there is considerable justification for the assumption that X_t is small and can be neglected, an assumption we make in this paper. In any case a significant X_t only slightly complicates the analysis and can be readily incorporated.

Another assumption made throughout most of the paper is that human capital is homogeneous in the sense that all units are perfect substitutes in production for each other and thus add the same amount to earnings. Of course, this assumption does not deny that some units may have been produced at considerably greater costs than others. The assumption of homogeneous human capital clearly differs in detail rather drastically from the usual emphasis on qualitative differences in education, training, and skills. I hope to demonstrate that these differences, while descriptively realistic and useful, are not required to understand the basic forces determining the distribution of earnings; indeed, they sometimes even distract attention from these determinants. Section 3*g* does, however, generalize the analysis to cover many kinds of human capital.

Chart 4 plots along the horizontal axis the amount invested in human capital measured for convenience by its cost rather than in physical units. Equal distances along the axis, therefore, do not necessarily measure equal numbers of physical units.

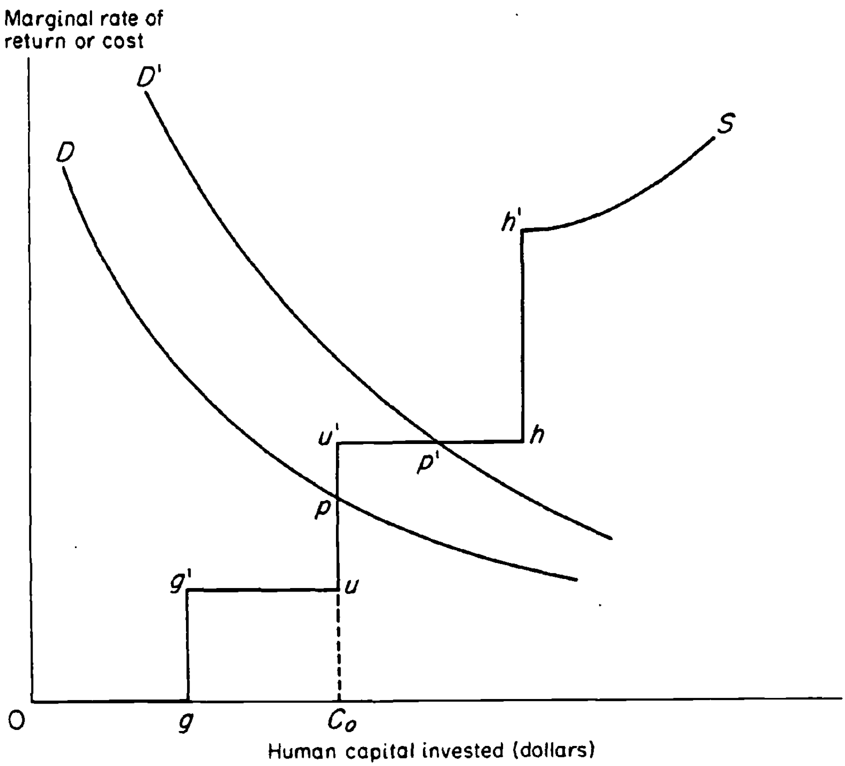
The curve D shows the marginal benefit, for simplicity measured by the rate of return, to a particular person on each additional dollar of investment, and is supposed to represent his demand curve for human capital. The curve S shows the effective marginal financing cost to him, measured for simplicity by the rate of interest, of each additional dollar invested, and represents in essence his supply curve of capital. If D exceeded S, the marginal rate of return would exceed the marginal rate of interest, and income would be increased by additional investment, while the opposite would be true if S exceeded D. Consequently, income is maximized by investing up to the point where $D = S$, given by p in the figure, and implying a total capital investment of OC_o.

b. Demand curves. The marginal rate of return depends on the time series of marginal returns and the marginal production cost of investment: if returns are constant for a long labor force period, it essentially equals the ratio of returns to these costs. Since all human

[79] See S. J. Mushkin, "Health as an Investment," *Journal of Political Economy*, 70, Special Supplement, October 1962, pp. 149–151.

CHART 4

Supply and Demand Curves for Investment in Human Capital



capital is assumed to be homogeneous, even an extremely large percentage change in the capital invested by any one person would have a negligible effect on the total quantity of capital available. Consequently, in order to explain why the demand curves for human capital in Chart 4 are negatively inclined and not horizontal, other effects of capital accumulation must be analyzed.

The principal characteristic that distinguishes human from other kinds of capital is that, by definition, the former is embedded or embodied in the person investing. This embodiment of human capital is the most important reason why marginal benefits decline as additional capital is accumulated. One obvious implication of embodiment is that since the memory capacity, physical size, etc. of each investor is limited, eventually diminishing returns set in from producing additional capital.[80] The result is increasing marginal costs of producing a dollar of returns.

Closely dependent on the embodiment of human capital is the importance of an investor's own time in the production of his own human capital.[81] Own time is so important that an increase in the amount invested in good part corresponds to an increase in the time spent investing:[82] in fact the only commonly used measures of schooling and training are years of schooling and training, measures entirely based on the input of own time. The cost of this time has been measured for several kinds of human capital, shown to be generally important, and given the name "foregone earnings." [83]

If the elasticities of substitution between own time and teachers, books, and other inputs were infinite, the use of own time and the deferral of investments could be avoided, without cost, aside from the limitations imposed by B, by an accumulation of all the desired capital instantaneously through complete substitution of other inputs for own time. If substitution were significantly imperfect (which is the more likely situation), the elimination of own time would cause the marginal costs of producing human capital to be higher and rise faster as capital was accumulated than if it was combined optimally with other inputs. In the latter case, however, the accumulation of capital is necessarily spread out over a period of calendar time called the

[80] If

$$h = f(I, B),$$

where h is the number of units of capital produced by a person per unit time, f is his production function, I is his capital investment in dollars per unit time, and B represents his physical and mental powers, then eventually

$$\frac{\partial^2 h}{\partial I^2} < 0.$$

[81] The production function in the previous footnote can be expanded to

$$h = f(R, T, B),$$

where R is the rate of input of other resources, and T is the rate of input of the investor's time per unit calendar time.

[82] If the horizontal axis in Chart 4 were replaced by one measuring investment time, the chart would be almost identical to those used in the "Austrian" theory of capital to explain optimal aging of trees or wine. Indeed, the main relevance of the Austrian approach in modern economics is to the study of investment in human capital!

[83] See T. W. Schultz, "Capital Formation by Education," *Journal of Political Economy*, 68, 1960, pp. 571–583.

"investment period." Presumably there are optimal combinations of inputs over an optimal investment period that maximizes the present value of benefits from a given capital investment. The spreading out of capital accumulation forced by the importance of own time can, however, only reduce but not eliminate the decline in marginal benefits as more is accumulated.[84]

In the first place, with finite lifetimes, later investments cannot produce returns for as long as earlier ones and, therefore, usually have smaller total benefits. This effect is important in societies with heavy adult mortality, but probably is not in the low mortality environment of modern Western societies. For unless fewer than approximately twenty years of working life remained, a reduction of, say, a year in the number of years remaining does not have much effect on the present value of benefits.[85] In the second place, later investments are less profitable than earlier ones because the present value of net benefits (or profits) is reduced merely by postponing them (and the reduction can be sizable, even for postponements of a few years).[86]

A third consideration is probably of great importance, although one cannot yet measure its quantitative significance. Since nobody can use his time at any activity without taking with him all of his human capital, the latter enters as an input along with his time in the production of additional capital. Initially, at young ages, the value of the time is small and probably even negative because parents or other baby-sitting services must be employed if he is not in school, or otherwise investing.[87] As he continues to invest, however, the capital accumulated becomes increasingly valuable, and so does his time.

Other things being the same, an increase in the value of time raises the marginal cost of later investments compared to earlier ones since the former use more expensive time. For any given rate of increase in its value as he ages, the costs of later investments are relatively greater, the larger the share of foregone earnings in costs and the smaller the

[84] The fact that a person's optimal stock of human capital is not immediately reached is often used in explaining the shape of his demand curve for human capital. On the problems in explaining why his optimal stock of nonhuman capital is not immediately reached, see D. W. Jorgenson, "The Theory of Investment Behavior," in *Determinants of Investment Behavior,* Universities-National Bureau Conference Series No. 18, Columbia University Press for NBER, 1967.

[85] For a demonstration of this, see pp. 47–48.

[86] See *ibid.,* pp. 72–73.

[87] For an attempt to measure the value of such services provided by elementary schools, see B. Weisbrod, "Education and Investment in Human Capital," *Journal of Political Economy,* 70, Special Supplement, October 1962, pp. 116–117.

elasticity of substitution between own time and other inputs.[88] One other thing that may not remain the same is the productivity of time: just as a greater amount of human capital is more productive than a lesser amount of capital in the rest of the economy, so too it may be more productive when used to produce additional human capital itself.[89] Marginal costs of later investments would not be greater if the increased productivity of own time was at least as great as its increased value. Because own human capital is carried along with own time, more productive or not, I am inclined to believe that its effect on productivity would be less, at least eventually, than its effect on the cost of own time. If so, the accumulation of human capital would on balance eventually increase later investment costs,[90] and thus decrease the present value of later benefits.

To digress a moment, the presumption that the marginal costs of typical firms are rising[91] is usually rationalized in terms of a limited "entrepreneurial capacity," an input that can only be imperfectly replaced by managers and other hired inputs. "Entrepreneurial capacity" is a construct developed to reconcile competition, linear homogeneous production functions, and determinate firm sizes, and most writers agree that there are no obvious empirical counterparts.[92] Indeed, the extremely large size achieved by many firms suggests that, frequently at least, entrepreneurial capacity is not very limiting. Per-

[88] This elasticity is relevant because investors may try to economize on their more costly time by substituting other inputs for time. Rough evidence of such substitution in education is found in the tendency for more valuable resources to be used per hour of the time of more advanced than less advanced students. The elasticity probably does not exceed unity, however, since the share of foregone earnings in total costs appears to rise with the level of education (see Schultz, *op. cit.*).

[89] If H measures the stock of human capital embodied in an investor, then the production function in footnote 81 can be expanded to include H, as in

$$h = f(R, T, H, B).$$

The productivity of greater human capital means a positive sign to $\partial h/\partial H$.

[90] Even if the effect on productivity continued to exceed that on the cost of own time, diminishing returns would cause the decrease in investment costs to become smaller and smaller over time. (For an illustration of this in a model that is quite similar to, although more rigorously developed than, the one presented here, see Y. Ben-Porath, "The Production of Human Capital and the Life Cycle of Earnings," *Journal of Political Economy*, August 1967). On the other hand, the decrease in the present value of benefits that results from a decrease in the number of years remaining would become larger and larger over time.

[91] This presumption can be justified by the observation that usually only firms producing a limited share of the output of an industry manage to survive.

[92] See M. Friedman, *Price Theory*, Chicago, 1962.

sons investing in human capital can be considered "firms" that combine such capital perhaps with other resources to produce earning power. Since "entrepreneurial" time is required to produce human capital, and since the latter is embodied in the entrepreneur, teachers, managers, and other hired resources can only imperfectly substitute for him. Therefore, in this case, "entrepreneurial capacity" is a definite concept, has a clear empirical counterpart, and, as has been indicated, can lead to significantly rising costs, which in turn limits the size of these "firms."

It is the sum of monetary benefits and the monetary equivalent of psychic benefits (which may be negative) from human capital, not just the former alone, that determines the demand curve for capital investment. If one makes the usual assumption of diminishing monetary equivalents, marginal psychic as well as monetary benefits would decline as capital is accumulated. The considerable uncertainty about future benefits also contributes to a negatively inclined demand curve if there is increasing marginal aversion to risk as more capital is accumulated.

c. Supply curves. The supply curves in Chart 4 show the marginal cost of financing, as opposed to producing, an additional unit of capital. The marginal cost of financing can be measured, for simplicity, by the rate of interest that must be paid to finance an additional dollar of capital. If the annual repayment required on a "loan" was constant for the remaining period of labor force participation, the marginal rate of interest would simply equal the annual repayment on an additional dollar of funds, adjusted upward for the finiteness of the labor force period.

If the capital market were homogeneous, with no segmentation due to special subsidies or taxes, transaction costs, legal restrictions on lending or borrowing, etc., and if risk were constant, even a large change in the amount of capital used by any person would have a negligible effect on his marginal cost of funds since it would have a negligible effect on the funds available to others. In the actual world, however, the market for human capital is extremely segmented: there are local subsidies to public elementary and high schools, state and federal subsidies to certain undergraduate and graduate students, transaction costs that often make own funds considerably cheaper than borrowed funds, and significant legal limitations on the kind of borrowing that is permitted. The result is that although certain sources of funds are cheaper than others, the amounts available to any person from the cheaper sources are usually rationed since the total demand for the funds tends to exceed their supply. This means that a person accumulating capital must shift from the cheapest to the second cheapest and on eventually to expensive sources. This shift from less to more expensive sources is primarily responsible for the positive inclination of the supply curve of funds even to one person. The rate of increase in each curve tends to be greater the greater the segmentation, since there is then greater diversity in the cost of different sources, with smaller amounts available from each.

The cheapest sources usually are gifts from parents, relatives, foundations, and governments that can be used only for investment in human capital. Their cost to investors is nil, and is represented in Chart 4 by the *Og* segment of the supply curve *S* that lies along the horizontal axis.[93]

Highly subsidized but not free loans from governments and universities, for example, that also can be used only for investment in human capital are somewhat more expensive: they are represented by the *g'u* segment of *S*. Then come the resources of investors themselves, including inheritances and other outright gifts, that could be used elsewhere. Their cost is measured by the foregone opportunities represented by the *u'h* segment of *S*. After these funds are exhausted, investors must turn either to commercial loans in the marketplace or to reductions in their own consumption during the investment period. These funds are usually available only at considerably higher, and somewhat rapidly rising costs: they are represented by the upward sloped segment *h'S*.

As emphasized earlier, the accumulation of human capital is not instantaneous, but is usually spread over a lengthy investment period. The rate of increase in financing costs, like that in production costs, would generally be less, the more slowly capital is accumulated because, for example, the accumulation of own resources could reduce the need to rely on more expensive sources.[94] The rate of increase in

[93] Conceptual separation of production costs from financial conditions suggests that direct government and private subsidies to educational institutions and other "firms" producing human capital might be included in the *Og* segment. When so separated, demand curves incorporate all production costs, not only those borne by investors themselves, supply curves incorporate all subsidies, and the rates of return relate "private" returns to "social" costs (for definitions of "private" and "social" see this volume, Chapter V).

[94] Superficially, there are many actual examples of the cost of funds depending on the period or stage of accumulation, such as the special subsidies to students of medicine or advanced physics. Many of these are best treated, however, as examples of a segmented capital market for different kinds of human capital, and are more appropriately discussed in section 3*g*, where the interaction among different kinds is analyzed.

each supply curve also depends, therefore, on the accumulation pattern that is chosen.

d. Equilibrium. Since both the stream of benefits and of financing costs depend on the path of capital accumulation, the latter cannot be chosen with respect to either alone. The rational decision is to select a path that maximizes the present value of "profits"; that is, the present value of the difference between these benefits and costs. With a model as general as the one presented so far, the supply and demand curves shown in Chart 4 would not be uniquely determined nor independent of each other. In order to justify, therefore, uniqueness and independence and to permit a relatively simple analysis of income distribution, it is sufficient to assume that own time and hired inputs are used in fixed proportions to produce human capital, that a unit of hired inputs is available at a given price, and that a unit of own time is also available at a given price (foregone earnings) up to a certain maximum amount, beyond which no time is available at any price. If the analysis of income distribution presented in this essay turns out to be useful, the implications of more general assumptions about the production of human capital should be explored.[95]

With these assumptions, the value of benefits is given by the area under the unique demand curve shown in Chart 4, the value of financing costs by that under the unique supply curve,[96] and the maximum difference is found by investing up to their point of intersection. At that point, marginal benefits equal marginal financing costs, which can be taken to mean that the marginal rate of return equals the marginal rate of interest.

Corresponding to the optimal accumulation path is an optimal investment period. If both the returns on each dollar invested and the repayments on each dollar borrowed were constant for the remaining labor force period, the current value of total profits, which is the difference between total returns and total repayments, would rise throughout the optimal investment period. A peak would be reached at the end, remain constant at that level throughout the labor force period, and then drop to zero.

95 A start is made by Ben-Porath, *op. cit.*, section 4.

96 For simplicity, the figures in this essay plot along the vertical axis marginal rates of return and interest on each additional dollar of investment rather than the present or current values of marginal benefits and financing costs. If returns and repayment costs were constant for indefinitely long periods, marginal rates of return and interest would exactly equal the current values of the flow of benefits and financing costs respectively on an additional dollar of investment.

The earnings actually measured in national income accounts do not purport to represent the profits on human capital. For one thing, the costs of funds are not deducted from returns, regardless of whether they consist of direct interest payments, foregone income, or undesired reductions in consumption. During the investment period, moreover, some and often all the costs of producing human capital are implicitly deducted before reporting earnings.[97] Consequently, measured earnings after the investment period only represent total returns, while during the period it is a hybrid of returns and production costs. I discuss first and most extensively the factors determining the distribution of measured earnings after the investment period, and only briefly consider the distribution of profits or of measured earnings during this period.

A major assumption of the remainder of this essay is that actual accumulation paths are always the same as optimal paths. Sufficient conditions for this assumption are that all persons are rational[98] and that neither uncertainty nor ignorance prevents them from achieving their aims. Of course, these are strong conditions, and a fuller model would make room for irrationality, uncertainty, discrepancies between actual and "desired" capital stocks, etc. Given, however, our rudimentary knowledge of the forces generating income distributions, it is instructive to determine how far even a simple model takes us. What impresses me about this model are the many insights it appears to provide into the forces generating inequality and skewness in the distribution of earnings and other income. In any case, it can be easily generalized to incorporate many of the considerations neglected, such as uncertainty, or discrepancies between actual and "desired" capital stocks.

3. The Distribution of Earnings

This model implies that the total amount invested in human capital differs among persons because of differences in either demand or supply conditions: those with higher demand or lower supply curves

[97] This intermingling of stocks and flows has many implications for age-earnings and age-wealth profiles that have been discussed elsewhere (see this volume, Chapters II, III, and VII).

[98] Since all persons are very young during much of their investment period, it may seem highly unrealistic to assume that their decisions are rational. Children have their decisions guided, however, as well as partly financed, by their parents, and as long as parents receive some monetary or psychic benefits from an increase in their children's economic well-being, parents have an incentive to help children make wise decisions.

invest more than others. There is some evidence that in the United States, persons with urban employment or high IQ and grades tend to invest more in formal education than those with rural employment or low IQ and grades partly because the former receive higher rates of return.[99] If the model is empirically correct, as assumed in the remainder of the essay, the sizable observed differences in education,[100] on-the-job training, and other kinds of human capital would suggest sizable differences in either one or both sets of curves.

Persons who invest relatively large amounts in themselves tend to receive relatively high profits and measured earnings after the investment period. If they invest more because of higher demand curves, as D' is higher than D in Chart 4, both the area under the demand curve and the difference between it and the area under a given supply curve is greater (compare point p' with p). If they invest more because of lower supply curves, the area under the supply curve for a given capital investment is smaller, and the difference between it and the area under a given demand curve, therefore, is greater.

a. "Egalitarian" approach. Instead of starting immediately with variations in both supply and demand conditions, I first treat a couple of important special cases. One of them assumes that demand conditions are the same for everyone, and that the only cause of inequality is differences in supply conditions. This can be considered an approximate representation of the "egalitarian" approach to the distributions of investments in human capital and earnings, which assumes that everyone more or less has the same capacity to benefit from investment in human capital. Investment and earnings differ because of differences in the environment; in luck, family wealth, subsidies, etc. which give some the opportunity to invest more than others. Eliminating environmental differences would eliminate these differences in opportunities, and thereby eliminate the important differences in earnings and investments.

Adam Smith took this view in his *The Wealth of Nations* when he said "The difference between the most dissimilar characters, between a philosopher and a common street porter, for example, seems to arise not so much from nature, as from habit, custom, and education." [101]

[99] See pp. 157–181.

[100] For example, the standard deviation of years of schooling exceeds three years in more countries.

[101] Modern Library edition, New York, 1937, p. 15. E. Cannan, ed., remarks that Smith was following David Hume, who said "consider how nearly equal all men are in their bodily force, and even in their mental powers and faculties, ere cultivated by education" (quoted *ibid.*).

Currently, many persons in the United States argue that most persons are intrinsically equally capable of benefiting from a college education; only poverty, ignorance, and prejudice prevent some from acquiring one.

Generally, the most important cause of differences in opportunities is differences in the availability of funds.[102] These in turn are derived from the same segmentation in the capital market which implies that cheaper funds are rationed, and that supply curves of funds are positively inclined even to individual investors. For a variety of reasons cheaper funds are more accessible to some persons than to others, and the former then have more favorable supply conditions. Some may live in areas providing liberal government and other subsidies to investment in human capital, or receive special scholarships because of luck or political contacts. Others may be born into wealthy families, have generous parents, borrow on favorable terms, or willingly forego consumption while investing. For all these reasons and more, supply curves of funds could differ considerably, and Chart 5 shows a few that differ in level or elasticity. For simplicity they are assumed to rise more continuously than the supply curve depicted in Chart 4.

If supply conditions alone varied, the equilibrium positions of different persons would be given by the intersections of the common demand curve with the different supply curves; the points p_1, p_2, p_3, and p_4 in Chart 5 represent a few such positions. Full knowledge of these positions, of the marginal rate of return associated with each amount of capital investment, would permit the common demand curve to be "identified." Moreover, the marginal rates could themselves be "identified" from the earnings received by persons with different capital investments.[103]

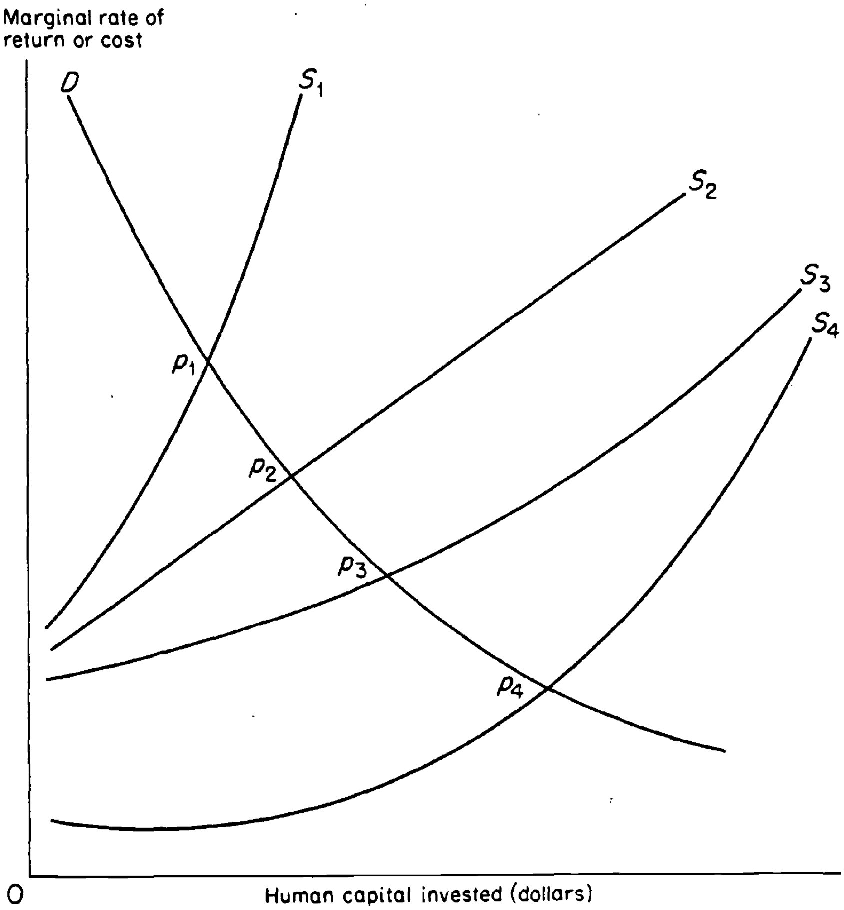
Persons with favorable supply conditions would invest relatively large amounts in themselves: the equilibrium positions in Chart 5 are further to the right, the lower the supply curves are. The distribution of the total capital invested obviously would be more unequal and skewed, the more unequal and skewed was the distribution of supply curves.

[102] Of course, it is not the only cause; for example, discrimination and nepotism are often important, and yet usually affect the benefits from rather than the financial costs of investing in human capital.

[103] Using the assumption that white males have the same demand curve for formal education, G. Hanoch first estimated the marginal rates of return to education from earning differentials between persons at different education levels, and then "identified" their common demand curve. See his *Personal Earnings and Investment in Schooling*, Ph.D. dissertation, University of Chicago, 1965, Chapter II.

CHART 5

Equilibrium Levels of Investment in Human Capital Resulting from Differences in "Opportunities"



If the labor force period was long, earnings would be related to the amount of capital invested by

$$E = \bar{r}C, \tag{80}$$

where E is earnings, C the total capital invested, and $\bar{r}$ the average rate of return on C. The distribution of E clearly depends on the distribution of C; indeed, if the demand curve for capital was completely elastic, $\bar{r}$ would be the same for everyone, and the distributions of earnings and investments would be identical except for a difference in units ($\bar{r}$) that depended on the aggregate supply of and demand for human capital. Since it is shown above that C is more unequally distributed and skewed the more unequal and skewed is the distribution of supply curves, the same applies to the distribution of E.

As we have seen, the demand curve for capital investment is usually negatively inclined rather than infinitely elastic primarily because human capital is embodied in investors. E will, therefore, usually be more equally distributed than C because a given percentage change in C will change E by a smaller percentage since $\bar{r}$ will decline as C increases and increase as C declines. Moreover, both E and C will be more unequally distributed and skewed the more elastic the demand curve is; for the greater the latter, the more that persons with favorable supply conditions would be encouraged to invest still more by a higher $\bar{r}$; and the more that those with unfavorable supply conditions would be encouraged to invest still less by a lower $\bar{r}$.

Similarly, an increase in the elasticities of supply curves that held constant their locations at the average value of C would also increase the inequality and skewness in E and C. Persons with unfavorable supply conditions would be encouraged to cut back their investments at the same time that those with favorable conditions were encouraged to expand theirs.

In the Mathematical Appendix exact relations between the distributions of E and C and the parameters of supply and demand curves are derived under the special assumption that all supply curves have the same constant elasticity, and that the demand curve also has a constant elasticity. Among the results of this more special model is that earnings are likely to be less unequally distributed and less skewed than supply curves (that is, than opportunities).[104]

b. "Elite" approach. At the other end of the spectrum is the assumption that supply conditions are identical and that demand conditions alone vary among persons. This can be considered an approximate representation of the "elite" approach to the distributions of investment in human capital and earnings, which assumes that everyone more or less has effectively equal opportunities. Actual investments and earnings differ primarily because of differences in the capacity to benefit from investment in human capital: some persons are abler and form an elite. In spite of the position taken by Smith and Hume, educational policy in England and some other parts of Europe has

[104] See section 6 of the Appendix.

been predicated on a version of the elite view: "There is a tendency of long historical standing in English educational thought (it is not nearly so visible in some other countries) to concentrate too much on the interests of the abler persons in any group that is being considered and to forget about the rest." [105]

Just as opportunities have been measured primarily by supply curves, so capacities are measured primarily by demand curves.[106] For a given (dollar) amount invested, persons with higher demand curves receive higher rates of return than others; or looked at differently, they have to invest more than others to lower the marginal rate to a given level. Since all human capital is assumed to be identical, demand curves can be higher only if more units of capital are produced by a given expenditure. It is natural to say that persons who produce more human capital from a given expenditure have more capacity or "ability." [107]

Since a higher demand curve means greater earnings from a given investment, in effect, ability is being measured indirectly; namely, by the earnings received when the investment in human capital is held constant.[108] This approach is an appealing compromise between defi-

[105] *Fifteen to Eighteen,* a report of the Central Advisory Council for Education, Geoffrey Crowther, Chairman, 1959, p. 87. In addition, many formal models of income distribution developed by economists are largely based on an underlying distribution of abilities. See, for example, A. D. Roy, "The Distribution of Earnings and of Individual Output," *Economic Journal,* 60, 1950, pp. 489–505, and B. Mandelbrot, "Paretian Distributions and Income Maximization," *Quarterly Journal of Economics,* 76, 1962, pp. 57–85.

[106] Let me repeat, however, that some differences in opportunities, such as those resulting from discrimination and nepotism, affect demand curves. Similarly, some differences in capacities affect supply curves.

[107] If the production function notation of footnote 81 is used, the i^{th} and j^{th} persons have the functions

$$h_i = f_i(R_i,\ T_i,\ B_i)$$
$$h_j = f_j(R_j,\ T_j,\ B_j).$$

The i^{th} person has more ability if $f_i > f_j$ when R and T, the inputs of market resources and own time, respectively, are held constant. If sometimes $f_i > f_j$ and sometimes $f_j > f_i$, there is no unique ranking of their abilities.

Note, however, that since demand curves incorporate psychic benefits and costs from human capital as well as monetary ones, i could have a higher demand curve than j, and thus be considered to have more capacity, simply because he receives more psychic benefits than j does.

[108] Note the similar definition by R. H. Tawney: "In so far as the individuals between whom comparison is made belong to a homogeneous group, whose members have equal opportunities of health and education, of entering remunerative occupations, and of obtaining access to profitable financial knowledge, it is plausible, no doubt, if all questions of chance and fortune are excluded, to treat the varying positions which they ultimately occupy as the expression of differences in their personal qualities" (*Equality,* Capricorn Books edition, New York, 1961, p. 121).

nitions of ability in terms of scores on IQ, personality, or motivation tests without regard to the effect on earnings, and definitions in terms of earnings without regard to opportunities.[109] The former pay excessive attention to form and not enough to results, while the latter hopelessly confound "nature" and "nurture," or ability and environment. Our approach directly relates ability to results, and at the same time recognizes the impact that environment has on results.[110]

If demand curves alone varied, the capital investments and marginal rates of return of different persons would be found at the intersections of the different demand curves with the common supply curve. In Chart 6 there clearly is a positive relation between the height of a demand curve, the amount of capital invested and the marginal rate. Knowledge of the latter two quantities for many different persons would permit an "identification" of the common supply curve, just as such information earlier permitted an "identification" of a common demand curve.

An important difference, however, is that the marginal rates themselves could not now be "identified" from information on the earnings and investments of different persons alone. In Chart 6 the marginal rate of return to investing OC_3 rather than OC_2 would be proportional to the area $p_2C_2C_3q_2$ for persons with the demand D_2 and to the larger area $q_3C_2C_3p_3$ for those with D_3. If a marginal rate was simply estimated from the difference in earnings between persons investing OC_2 and OC_3, the estimate would be proportional to $D_2p_2C_2C_3p_3D_3$, which clearly greatly exceeds both true rates. To arrive at correct estimates, either the earnings of persons investing OC_2 would be adjusted upward by the area $D_3D_2p_2q_3$, or the earnings of those investing OC_3 adjusted downward by the area $D_3D_2q_2p_3$.[111] Note, incidentally, that those arguing that most of the differences in earnings between persons at different levels of education or training result from differences in ability are essentially assuming a common supply curve and steeply inclined demand curves.

Aside from chance, Tawney mainly stresses the importance of holding constant health, education, and financial knowledge, which are simply different kinds of human capital.

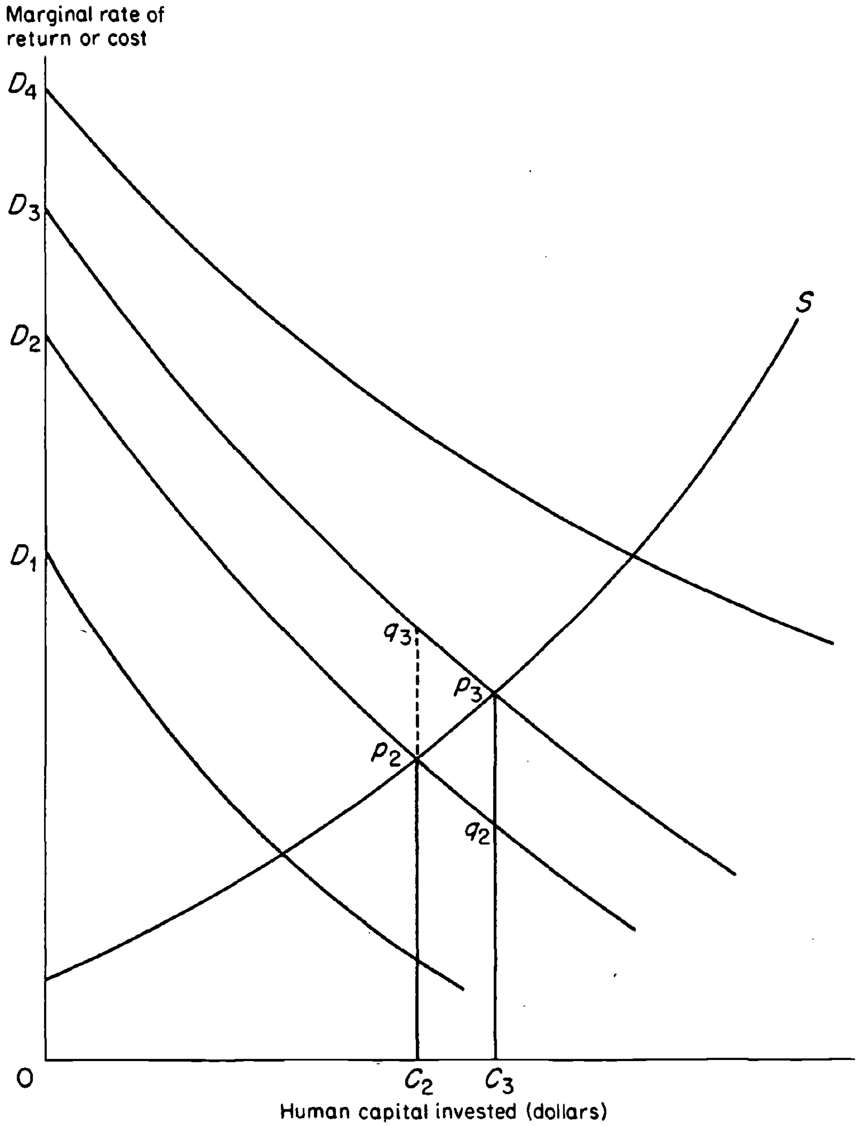
[109] For a review of these definitions in the context of analyzing income distributions, see H. Staehle, "Ability, Wages, and Income," *The Review of Economics and Statistics,* 25, 1943, pp. 77–87.

[110] I have not tried to explain why some people are "abler" than others; this might ultimately be traced back to differences in numerous basic ability "factors." For a model of this kind, see Mandelbrot, *op. cit.*

[111] Some adjustments along these lines to estimated rates of return on formal education can be found in this volume, pp. 202–204.

CHART 6

Equilibrium Levels of Investment in Human Capital Resulting from Differences in "Abilities"



Earnings and capital investments are clearly more unequally distributed and skewed the more unequally distributed and skewed are demand curves. The same kind of arguments as those used in the previous section should make it apparent that both distributions are also more unequal and skewed, the greater the elasticities of the supply and demand curves. If the supply curve was positively inclined, the average rate of return would tend to be greater, the larger the amount invested. Therefore, earnings would tend to be more unequally distributed and skewed than investments.

In the Mathematical Appendix exact relations between the distributions of earnings and capital investment and the parameters of the supply and demand curves are derived under the special assumption that all demand curves have the same constant elasticity, and that the supply curve also has a constant elasticity. One of the more interesting results is that earnings and investments would *necessarily* be more unequally distributed and skewed than demand curves.[112] If, for example, demand curves (i.e., capacities) were symmetrically distributed, both earnings and investments would be skewed to the right.

c. A comparison of these approaches. Before moving on to the general case that incorporates variations in both supply and demand conditions, it is illuminating to contrast the more important implications of these special cases. For under the guise of the "egalitarian" and "elite" approaches, they are frequently explicitly advanced and are still more widely implicitly assumed.

The "egalitarian" approach implies that the marginal rate of return is lower, the larger the amount invested in human capital, while the "elite" approach implies the opposite relation. Marginal rates of return appear[113] to decline in the United States as years of schooling increase, which supports the "egalitarian" approach. However, in Canada, a country in many economic respects quite similar to the United States, estimated marginal rates do not decline consistently as schooling increases.[114]

[112] See section 4 of the mathematical appendix on p. 138.

[113] I say "appear" because these rates have not been fully corrected for differences in the average level of "ability" at different education levels; such a correction might eliminate the apparent decline (see Hanoch, *op. cit.*, or this volume, p. 202).

[114] See J. R. Podoluk, *Earnings and Education,* Dominion Bureau of Statistics, 1965. Note that since different years of schooling are not perfect substitutes, the pattern of rates are also affected by the relative demand for and supply of different years. Thus the relatively small number of college-educated persons in Canada might explain the relatively high rates of return to college education there.

The inequality in earnings tends to be less than that in supply conditions in the "egalitarian" approach, and greater than that in demand conditions in the "elite" approach because the former implies a negative, and the latter a positive, correlation between rates of return and amounts invested. Put differently and perhaps more interestingly, to understand the observed inequality in earnings, the "egalitarian" approach has to presume greater inequality in opportunities than the "elite" one has to presume about capacities. Inequality in earnings is a more serious problem to the former, therefore, in the sense that a given observed amount implies greater underlying "inequities" or "noncompeting groups" [115] than it does to the latter.

For a similar reason, the positive skewness in earnings is probably less than that in opportunities under the "egalitarian" approach and greater than that in capacities under the "elite" approach. Indeed, as pointed out in the last section, it is shown in the Mathematical Appendix using the assumptions of constant and identical elasticities of demand, and a constant elasticity of supply, that a symmetrical distribution of capacities *necessarily* results in a positively skewed distribution of earnings. Therefore, an age-old problem of economists—how to reconcile a skewed distribution of income with a presumed symmetrical normal distribution of abilities[116]—turns out to be no problem at all.[117] In the "egalitarian" approach, on the other hand, observed skewness is more difficult to explain because it implies still greater skewness in the distribution of opportunities, a skewness that may be associated with a skewed distribution of gifts and inheritance, etc.[118]

[115] The interpretation of income inequality in terms of noncompeting groups was popular among nineteenth and early twentieth century writers. For a review see H. Dalton, *Some Aspects of the Inequality of Incomes in Modern Communities*, London, 1920, Part II. "Groups" may be noncompeting either because of differences in opportunities, as assumed in the "egalitarian" approach, or because of differences in capacities, as assumed in the "elite" approach.

[116] For example, A. C. Pigou said "Now, on the face of things, we should expect that, if as there is reason to think, people's capacities are distributed on a plan of this kind [i.e., according to a symmetrical normal distribution], their incomes will be distributed in the same way. Why is not this expectation realized?" *The Economics of Welfare*, 4th edition, New York, 1950, p. 650. See also P. A. Samuelson, *Economics*, 6th edition, New York, 1964, pp. 120–121.

[117] It is not possible, however, to reconcile extremely large skewness in earnings with a symmetrical distribution of capacities.

[118] Pigou's principal answer to the question he sets out in footnote 116 is largely based on a presumed skewed distribution of inheritances, which affects, among other things, the distribution of investments in training (*ibid.*, pp. 651–654). Or, as Allyn Young said, "The worst thing in the present situation is undoubtedly the

d. A more general approach. If either all demand or all supply curves were identical, the supply and demand curves of persons investing the same amount would also be identical if different demand or different supply curves did not touch in the relevant region. This, in turn, means that all persons investing the same amount would have identical earnings. Yet if the amount invested is measured by years of schooling, there is abundant evidence of considerable variability in the earnings of persons with the same investment.[119] Possibly improved measures of investment or the introduction of transitory earnings would eliminate most of the variability; I suspect, however, that a significant portion would remain. If so, neither special case—that is neither variations in demand nor in supply curves alone—is sufficient, although one set of curves might vary much more than the other.

If both supply and demand curves varied, different persons could invest the same amount, and yet some could earn more than others because they had higher demand (*and* supply) curves; in Chart 7, the same amount would be invested by persons with D_3 and S_1, D_2 and S_2, and D_1 and S_3. As this example indicates, knowledge of the various equilibrium marginal rates of return and investments would no longer be sufficient to identify either a supply or a demand curve because the equilibrium positions would be on different curves. Moreover, again the marginal rates themselves could not be identified from information on earnings and investments alone because persons with different investments would generally have different demand curves.

The distributions of earnings and investments would partly depend on the same parameters already discussed: both would be more unequal and skewed, the greater the elasticities of supply and demand curves, and the more unequal and skewed their distributions. The distributions of earnings and investments also depend, how-

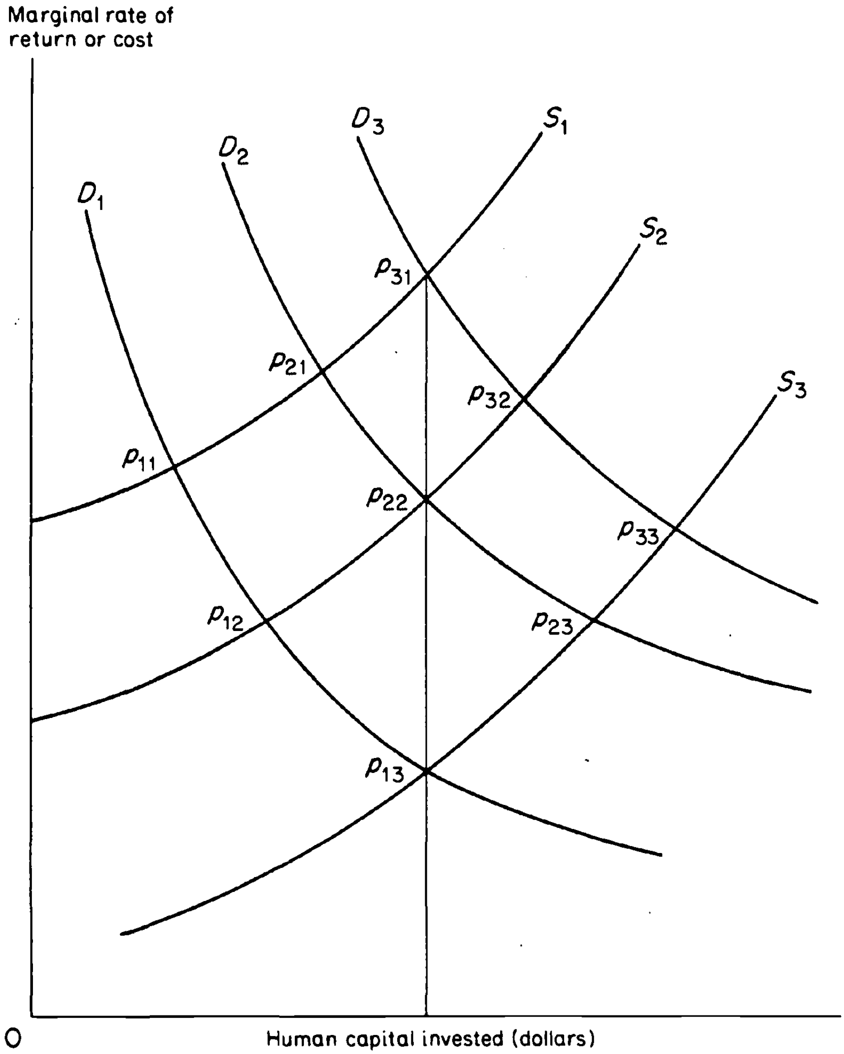
---

extreme skewness of the income frequency curve . . . reflecting as it undoubtedly does the presence of a high degree of inequality in the distribution of opportunity" ("Do the Statistics of the Concentration of Wealth in the United States Mean What they are Commonly Assumed to Mean?," *Journal of the American Statistical Association*, 15, 1917, pp. 481–482). One should point out, however, that even "a high degree of inequality in the distribution of opportunity" is not sufficient to produce skewness in earnings, and that skewed distribution of opportunities is necessary, at least in the "egalitarian" approach.

119 For example, the coefficient of variation in the incomes of white males aged 35–44 in 1949 was 0.60 for high school graduates and 0.75 for college graduates (see this volume, Table 9, p. 182). Or in 1959, years of schooling explained less than 20 per cent of the variance in the earnings of white males aged 25–64 in both the South and non-South (see Table 1, p. 92).

CHART 7

Equilibrium Levels of Investment in Human Capital Resulting from Differences in "Abilities" and "Opportunities"



ever, on a new parameter: namely, the correlation between different curves.

There are several reasons why supply conditions do not vary independently of demand conditions. Abler persons are more likely to receive public and private scholarships, and thus have their supply curves shifted downward. Or children from higher-income families probably, on the average, are more intelligent and receive greater psychic benefits from human capital. On the other hand, private and public "wars" on poverty can significantly lower the supply curves of some poor persons. Since the first two considerations have, unquestionably, been stronger than the third, it is reasonable to presume a positive[120] correlation between supply and demand conditions, perhaps a sizable one.

If supply and demand curves were uncorrelated, one might have the equilibrium positions given by p_{31}, p_{32}, and p_{33} in Chart 7; if they were negatively correlated, by p_{31}, p_{22}, and p_{13}; and if they were positively correlated, by p_{11}, p_{22}, and p_{33}. The chart clearly shows that a positive correlation increases the inequality in both investments and earnings; it also increases skewness by increasing the earnings and investments of persons who would have relatively high earnings and investments anyway.

An impression of a negative correlation between supply and demand conditions—that is, between opportunities and capacities—is sometimes obtained from persons investing the same amount. As the curves D_3 and S_1, D_2 and S_2, and D_1 and S_3 in Chart 7 clearly show, however, the supply and demand curves of persons investing the same amount *must* be negatively correlated, regardless of the true overall correlation between them. Valid evidence of this latter correlation is provided by information on the amount of variation in earnings "explained" (in the analysis of variance sense) by the variation in investments. For example, if the correlation between supply and demand curves was perfect and positive, *all* the variation in earnings would be "explained" by investments. Moreover, the smaller the algebraic value of this correlation, the less the variation in earnings is "explained" by investments, and the more that earnings vary among persons making the same investment.

Supplement: Estimating the Effect of Family Background on Earnings[121]

One important implication of the above analysis on the interaction between opportunities and capacities (i.e., supply and demand condi-

120 By "positive" is meant that more favorable demand conditions are associated with more favorable supply conditions.

121 The issues considered in this addendum were already briefly considered by J. Mincer in "The Distribution of Labor Incomes: A Survey," *Journal of Economic Literature,* March 1970, p. 20.

tions) was mentioned but not sufficiently stressed—namely, that opportunities and capacities would be negatively correlated for persons investing the same amount, regardless of the overall correlation between them. For example, if, as many studies suggest, investment opportunities are less favorable to children in large families, ability and number of siblings would be *positively* correlated for persons investing the same amount, even if they were negatively correlated in the population as a whole.[122]

I want to stress, however, its importance in separating the effect of family background on earnings from the effect of schooling. In recent years, many studies have tried to separate these effects by running multiple regressions of earnings on years of schooling and background variables (and often other variables as well), and using the schooling regression coefficients as a measure of the independent effect of schooling, and the background coefficient as a measure of the independent effect of background.[123] Yet if years of schooling result not from random behavior but from optimizing behavior, these studies may be seriously understating the contribution of background to earnings and overstating that of schooling.

To show this in a dramatic fashion, assume that family background only has an effect on opportunities, capacities being independent of background, and that background is the only variable affecting opportunities. A series of equilibrium positions are shown in Chart 8: b_1, b_2, b_3, and b_4 are supply curves of particular persons with different background (b_4 is a more favorable background than b_3, etcetera), and d_1, d_2, d_3, and d_4 are their corresponding demand curves (or capacities). The optimal accumulation of capital by each person is given by the intersection of his supply and demand curve, or by the points e_1, e_2, e_3, and e_4.

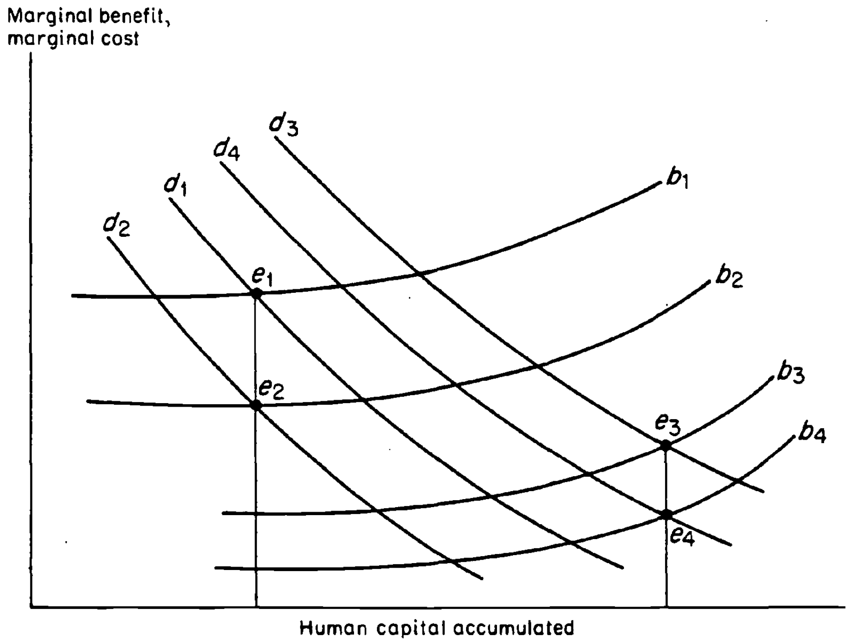
If the independent effect of background on earnings were estimated from differences in earnings between persons with the same accumulation of human capital but different backgrounds, point e_2 would be compared to e_1 or point e_4 to e_3. In both comparisons, persons with

[122] This implication was derived and tested many years ago by R. A. Fisher. See his *The Genetical Theory of Natural Selection*, 2nd ed., New York, 1958, Chapter 11.

[123] See, for example, S. Bowles, "Schooling and Inequality from Generation to Generation," *Journal of Political Economy*, 80, 3, Supplement, May–June, 1972; A. Leibowitz, "Home Investments in Children," *Journal of Political Economy*, 82, 2, Supplement, March–April, 1974; J. Coleman and P. Rossi, "Processes of Change in Occupation and Income," mimeograph, 1974; or Louis Levy-Garboua, "Does Schooling Pay?," *Entroit de Consommation*, 3, 1973.

CHART 8

The Effect of Background and Ability on Earnings and the Accumulation of Human Capital



higher backgrounds have *lower* earnings,[124] yet in this example a better background certainly does not lower earning capacity. Of course, the reason for this erroneous result is that persons with superior backgrounds (i.e., superior opportunities) would accumulate the same amount of human capital as those with inferior backgrounds (i.e., inferior opportunities) only if the former have inferior earning capacities.

A similar argument shows that the effect of human capital on earnings is overstated when estimated from differences in earnings between persons with the same background but different accumulations of human capital. For they choose to invest different amounts only be-

124 Leibowitz, *op. cit.*, Tables 4–6, does find a negative effect of mother's education on earnings when own years of schooling are held constant. (However, she also finds a positive coefficient for parent's income.)

cause they differ in earning capacity.[125] In other words, the effect of human capital on earnings would be more accurately estimated if background were omitted as a separate variable than if held constant.

The argument holds, more generally, when background affects earning capacity as well as opportunities, and other variables affect opportunities as well as capacities, if background is a major determinant of opportunities, which is surely true, and other variables are a major determinant of capacities, which is also true. A multiple regression of earnings on background and human capital would then always understate the effect of background and overstate that of human capital.

In principle, the appropriate statistical procedure is a simultaneous equations model that would "identify" the opportunities and capacities functions, including the effects on both functions of background and human capital accumulation. In practice, however, a good specification is not easily obtained because information on variables that can "shift" these functions is limited. Some progress can be made, however, by starting with simple models. For example, the model depicted in Chart 8 can be written as

$$\begin{aligned} r_d &= a + bH + u \\ r_s &= \alpha + \partial H + \gamma B + v \\ r_d &= r_s, \end{aligned}$$

where r_d is the marginal rate of return on, and r_s is the marginal cost of, financing, the human capital accumulation H, u, and v represent disturbance terms that are uncorrelated with each other and with the exogenous background variable B, and $b < 0$, $\partial > 0$, and $\gamma < 0$. Then the demand function can be identified—i.e., b can be estimated—by using the reduced form equations: r regressed on B and H regressed on B. The supply function is not identifiable in this system but would be also if there were information on exogenous "ability" variables that only entered the demand function.

e. The effects of age. A common method of explaining the rise in the inequality of earnings with age is to introduce random influences and let their effects partly accumulate over time.[126] An alternative method suggested by our analysis is to introduce earnings during the investment period. It has already been stressed that investing in hu-

125 Again, Leibowitz finds that controlling for background tends to *raise* the estimated effect of schooling.

126 See, for example, J. Aitchison and J. A. C. Brown, *The Lognormal Distribution,* London, 1957, pp. 108–111.

man capital takes time primarily because an investor's own time is an important input into the investment process. Persons who invest relatively little tend also to cease investing at relatively early ages; for example, dropouts from elementary school generally cease investing before college graduates do. If, therefore, persons with higher demand or lower supply curves tend to have longer investment periods as well as larger investments,[127] the accrued earnings (that is, the area under demand curves) of persons with high earnings would increase for longer periods than would those of others. The effect would be greater inequality in the distribution of accrued earnings at older ages as long as an appreciable number of persons are still investing. In other words, inequality could rise with age because it takes abler persons and those with favorable opportunities longer to reach their full earning power.

Measured earnings differ, however, from accrued earnings during investment periods partly because the income and capital accounts are confounded: measured are derived from accrued earnings only after some investment costs are deducted. Since the amounts deducted are large and variable, measured earnings during investment periods may be only weakly positively or even negatively correlated with earnings afterwards.[128] The effect of mixing earnings and investment costs on the inequality in earnings is less clear-cut. On the one hand, inequality is decreased because high earners invest larger amounts and for longer periods; on the other hand, the inequality is increased by the variation among persons in the amounts deducted.

f. Profits on human capital. The profits on investments in human capital are not measured by earnings or returns alone, but rather by the difference between returns and repayment costs. Geometrically, they do not equal the area under a demand curve alone, but the difference between the areas under a demand and supply curve. Although profits are obviously less than earnings, the former are not necessarily distributed either less or more unequally than the latter. Indeed, a very useful theorem proved in the Appendix states that if all demand curves had the same constant elasticity, and if supply curves also did, the percentage difference between earnings and profits would be the same for everyone, and thus the distribution of profits would be exactly

127 They necessarily have larger investments if supply and demand curves are not negatively correlated.

128 For one piece of evidence indicating virtually no correlation, see J. Mincer, "On-the-Job Training: Costs, Returns, and Some Implications," *Journal of Political Economy*, 70, Special Supplement, October 1962, p. 53.

the same, aside from scale, as the distribution of earnings.[129] If constant and identical elasticities can be taken as a rough first approximation to the truth, there is no need to dwell on the distribution of profits, for it would depend on exactly the same variables already discussed in detail for earnings. To summarize and apply those results: profits would be more unequally distributed and skewed, the more unequally distributed and skewed were the supply and demand curves, the greater the positive correlation between these curves, and the greater their elasticities.

g. Heterogeneous human capital. A major assumption has been that all human capital is homogeneous, an assumption that conflicts with obvious qualitative differences in types of education, on-the-job training, informal learning, etc. in the same way that the frequently used assumption of homogeneous physical capital conflicts with myriad observed differences in plant, equipment, etc. The advantage of these assumptions is that by sweeping away qualitative detail—detail that, incidentally, has received excessive attention in the literature on human capital—one can concentrate on more fundamental relationships.

For those unable to accept, even tentatively, an assumption of homogeneous human capital, let me hasten to stress that different kinds can rather easily be incorporated into the analysis. For example, with two kinds of capital, each person would have two sets of demand and supply curves, and in equilibrium, marginal benefits and financing costs would be equal for each set. The distribution of earnings would still depend in the same way on the distributions and elasticities of the supply and demand curves. The only significant new parameters introduced are those giving the correlations between the different supply and also between the different demand curves for the two kinds of capital. These correlations measure the extent to which people are relatively able or have access to funds on relatively favorable terms for both kinds. These correlations are presumably positive, but by no means perfect, because both ability and access to funds carry over to some extent, but not perfectly, from one kind of capital to another. It should be intuitively clear that positive correlations tend to make both earnings and investments more unequally distributed and skewed, for then persons who invest much (or little) in and earn much (or little) from one kind of capital also tend to invest and earn much (or little) from the other.

129 See section 8 of the Mathematical Appendix for a proof (pp. 143–145).

4. Some Applications

The supply and demand curves for investment in human capital that affect the distribution of earnings are not immutable, but are partly determined by aspirations, private generosity, and public policy. Although in a long run perspective all may be subject to change, the variables influencing opportunities are more easily and immediately influenced than are those influencing capacities. It is not surprising, therefore, that the various institutions discussed in this section all influence the distribution of earnings through their impact on the distribution of opportunities. The institutions covered are rather diverse: more "equal" opportunity, admission to education and other training institutions on the basis of "objective" testing, compulsory minimum schooling and other investments, and improvements in the capital market. Their diversity and importance demonstrates the value of our analysis in relating the distribution of earnings to private and public actions.

a. Equality of opportunity. An avowed goal of many countries, especially the United States, has been to achieve "equality of opportunity," yet the meaning and implications of this term have not been carefully explored.[130] A natural statement within our framework is: equality of opportunity implies that all supply curves are identical, with opportunity being more unequal, the greater their dispersion. A full definition might also include elimination of public and private discrimination and nepotism, which would limit the dispersion in demand curves as well. The implications of this statement can be easily derived by building on the analysis developed in the last section.

For example, if supply curves were identical and discrimination and nepotism eliminated, earnings and investments would differ essentially because of differences in capacities, a major goal of those advocating equal opportunity.[131] Therefore, the "equalitarian" ap-

[130] One of the better statements is by Tawney: "[Equality of opportunity] obtains in so far as, and only in so far as, each member of a community, whatever his birth, or occupation, or social position, possesses in fact, and not merely in form, equal chances of using to the full his natural endowments of physique, of character, and of intelligence" (*Equality, op. cit.*, p. 106). Tawney does not, however, relate this definition to any analysis of income inequality.

[131] Tawney said: "The inequality which they [and Tawney] deplore is not inequality of personal gifts, but of the social and economic environment" (*Ibid.*, p. 38). Inequality in position in Michael Young's "meritocracy" is entirely related to inequality in ability. See his *The Rise of the Meritocracy 1870–2033*, New York, 1959.

proach to distribution implies that equalizing opportunity would essentially eliminate all the inequality in earnings and investments, while the "elite" approach denies that it would make any essential difference.

The effect of equal opportunity on the inequality in earnings and investments is also clear-cut. If supply and demand curves were not negatively correlated, and if equal opportunity did not raise the algebraic value of this correlation or the absolute values of the supply and demand elasticities, then it must reduce the inequality in earnings. The reason is simply that one of the basic determinants of inequality, the dispersion in supply curves (and partly also in demand curves), is eliminated while the others are not affected "perversely." Unless the correlation between supply and demand curves was positive and sizable, the reduction in the inequality of earnings would be less than that in investments, however, because the elimination of inequality in supply curves would raise the correlation between investments and rates of return, which would partly offset the reduction in the inequality of investments.

Identical supply curves can be achieved in many ways: subsidies to institutions providing investments, such as through the public schools; scholarships to investors, especially poorer ones; government-financed or insured loans to investors; "head start" programs for poorer children; and so on. Probably the most desirable system is to subsidize the external or "neighborhood" cultural, political, and economic benefits of investments, and develop loan programs to finance the direct or private benefits.[132] Only if external benefits completely dominated, which is not true even for education,[133] would this lead to "free" investments.

b. Objective selection. Often confused with policies that equalize opportunities are those that ration entrance into highly subsidized schools and other investment institutions not by "favoritism," but by "objective" standards, such as scores on special examinations, grades or class standing in prior training, etc. Of the many examples throughout

[132] The first and most discussed proposal along these lines can be found in M. Friedman, "The Role of Government in Education," in *Economics and the Public Interest,* Robert A. Solow, ed., New Brunswick, N.J., 1955, pp. 123–144. For a variant of the Friedman proposal see W. Vickrey, "A Proposal for Student Loans," in *Economics of Higher Education,* S. J. Mushkin, ed., Washington, D.C., 1962, pp. 268–280.

[133] See pp. 196–198.

the world, one can mention the examinations at the end of middle schooling in Japan that have sharply limited the number going on to public higher schools,[134] the "eleven plus" examinations in Great Britain that have determined entrance into grammar schools,[135] or the high-school records that help determine admission to the University of California, Harvard University, and many other universities in the United States.

"Objective" selection is an illusion in the "egalitarian" approach to distribution because if differences in capacities are unimportant, selection cannot be "objective" and must be subjective. This explains why there is continuous pressure on public universities in the United States to admit essentially all applicants meeting minimum qualifications. There is greater justification within the "elite" approach; not surprisingly, therefore, "objective" selection has been prominent in countries like Great Britain and Japan. If differences in capacities are obvious and substantial, tests administered even as early as ages eleven or fifteen can select the more promising students.

Generally, persons failing examinations or other standards are not prevented from continuing their investments; only the cost of funds to them is greater. Objective standards clearly do not, therefore, equalize opportunity because persons selected obtain funds under relatively favorable conditions. Since a system of objective standards, if used successfully, also tends to increase the positive correlation between supply and demand conditions, the resulting inequality in earnings and investments would exceed that under equality of opportunity.[136] Indeed, the resulting inequality would even tend to exceed that under a system selecting applicants at random because objective standards encourage abler persons, who probably earn and invest more than others anyway, to earn and invest still more because they are heavily subsidized.[137]

[134] The pressure felt by parents and pupils has been known as "the examination hell." See H. Passin, *Society and Education in Japan,* Teachers College, Columbia University, 1965, pp. 104–108.

[135] About four-fifths of the eleven- and twelve-year-olds are not admitted to grammar schools. See Crowther, *op. cit.,* p. 87. Although the others generally can continue their education in "all age" or "secondary modern" schools, their chances of continuing beyond age fifteen are considerably reduced because many of these schools do not provide fifth and higher years of secondary schooling (*ibid.*).

[136] Assuming, of course, that the elasticities of supply and demand curves and the distribution of the latter are the same in both situations.

[137] In the quote from the Crowther report in section 3*b* above there is an implied criticism of the system of objective selection in Great Britain.

c. Compulsory minimum investments. Virtually every country has laws requiring a minimum investment in human capital. Usually only a minimum number of school years is required,[138] although sometimes apprenticeship programs and other on-the-job training can be substituted. Since differences in the generosity or wealth of parents are a major cause of inequality in the "egalitarian" approach to income distribution, minimum investment legislation would make sense under that approach. For by imposing minimum standards, poorer and less generous parents are forced to spend more on their children, which improves the latter's opportunities and earnings. Since differences in capacities are the major cause of inequality in the "elite" approach, there would not be much interest in these standards under that approach.[139]

In Chart 9 let D be the demand curve and S the supply curve of a particular person in the absence of minimum standards, and OC_e be his equilibrium capital investment. If legislation is passed requiring at least OC_m, his supply curve would be shifted to the curve C_mS because his parents are forced to become more "generous," and his equilibrium investment would be increased to OC_m. The distribution of investments would be truncated at OC_m, with everyone who would have invested less being brought up to that point.[140]

Truncating the distribution of investments reduces the inequality in investments and through that in earnings as well. In effect, the inequality of opportunity is reduced by bringing supply curves closer together;[141] indeed, by setting the minimum standard at least equal

138 For example, the British in practice require only attendance through age fifteen. The Act of 1944 also required part-time education for those leaving before age sixteen, but it has not been put into effect (see Crowther, *op. cit.*, p. 105); for the first half of this century the Dutch simply required seven consecutive years of schooling. See M. M. Loren, *Education in the Netherlands*, Netherlands Information Bureau, New York, 1942, p. 12.

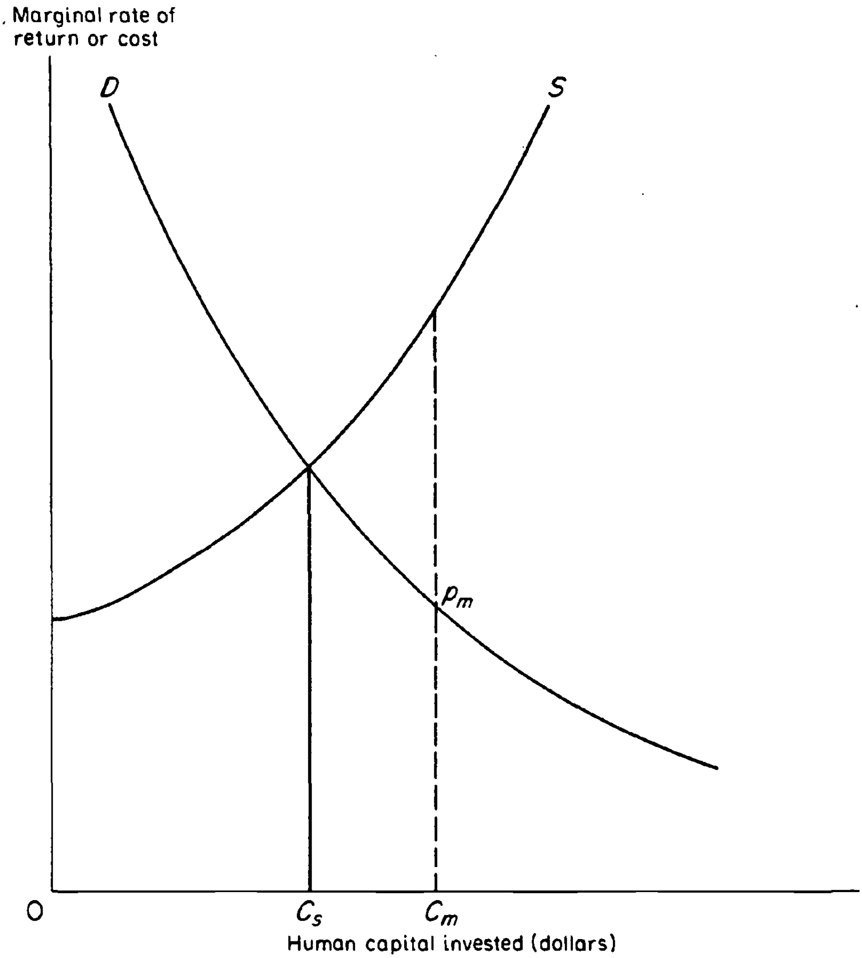
139 Except, of course, to the extent that investment in human capital produces significant external benefits.

140 In 1900 the Dutch passed a law providing for at least seven years of schooling; by 1960 virtually no one in the male labor force had less and almost 60 per cent were about at that level. Similarly, in 1951 about two-thirds of the persons in the labor force in Great Britain had nine years of schooling, the minimum amount required, and only one in ten had less. See B. R. Chiswick, *Human Capital and the Distribution of Personal Income,* Ph.D. dissertation, Columbia University, 1967. Educational distributions are not so truncated in the United States because different states passed laws at different times and heavy internal migration and immigration from abroad moved people from their educational origins.

141 If minimum standards, apply only to one kind of human capital, such as schooling, some parental funds may simply be drawn away from other kinds, such as on-the-job training, which would increase the dispersion in the supply of funds to these kinds.

CHART 9

The Effect of a Compulsory Investment Law on the Amount Invested



to the maximum amount anyone would have invested, identical investments and full equality of opportunity could be obtained. Moreover, opportunity is equalized at the same time that the elasticities of supply are drastically reduced (compare S with C_mS in Chart 9), which also contributes to a reduction in the inequality of earnings and investments. As section 4e shows, however, greater equality is obtained only at the expense of a less efficient allocation of the total investment in human capital.

Since the purpose of minimum standards is to offset the effects of poverty and niggardliness, appropriate subsidies could in principle achieve the same result without compulsion. The effectiveness of voluntary investment in human capital is often underrated [142] because subsidies to human capital usually cover, at best, only a portion of the earnings foregone. If they covered all costs, including those foregone, almost all children, I am confident, would continue in school through the age desired.

d. Improvements in the capital market. It was emphasized earlier that the major cause of both the rise in the cost of funds as a person's investments increased and the differences in the cost to different persons is the rationing of cheaper sources of funds due to a segmentation of the capital market. Government funds are generally the cheapest because of subsidies, own funds are usually cheaper than those borrowed commercially because of transaction costs and the difficulty of using human capital as collateral, and so forth. Less segmentation—due, for example, to a reduction in subsidies or transaction costs—would narrow the spread among different sources, and thus both increase the elasticities of supply curves and reduce the dispersion among them.[143]

An improvement in the capital market would, therefore, have somewhat conflicting effects on the distributions of earnings and of investments. A narrower dispersion of supply conditions reduces the inequality, while increased elasticities of supply increase both the inequality and skewness in earnings and investment. Apparently, skewness would tend to increase, but inequality could go in either direction.

e. Equality and efficiency. The discussion of capital market improvements, minimum investment legislation, and other changes in opportunities has emphasized inequality and skewness and ignored efficiency. Yet the trade off and conflict between equity and efficiency have occupied as much time of social commentators as any other economic issue. One important advantage of the analytical framework developed in this study is, therefore, that efficiency can be as easily and

142 "Voluntary staying-on seems both too haphazard and too precarious to be depended on as the basis of a national system" (Crowther, *op. cit.*, p. 107).

143 The reduced segmentation would contribute, therefore, to greater equality of opportunity.

systematically handled as distribution. This is now illustrated with some examples previously discussed. It is shown, in particular, that, while equality and efficiency are sometimes affected differently, they also sometimes change in the same direction.

If subjective attitudes toward risk are ignored, the criterion for an efficient allocation of the total investment in human capital is well known; namely, that the marginal social rate of return be the same for all persons. If one assumes that the ratio of social to private rates is identical for everyone, this criterion simply requires equality between all marginal private rates, while inefficiency can be measured quantitatively essentially by the inequality in these rates. Efficient allocation of investment between human and other capital requires that the marginal rates on each type of capital be equal to each other, but this aspect of efficiency is not discussed here.

If all supply curves have infinite elasticities, equalizing opportunities not only reduces the inequality in earnings and investments, but also reduces the inequality in marginal rates, which means that the total investment in human capital is allocated more efficiently. A reduction in the segmentation of the capital market that increases the elasticities and reduces the dispersion of supply curves may or may not reduce the inequality in earnings and investments, but does tend to reduce the inequality in marginal rates, and thus improve efficiency. On the other hand, compulsory minimum standards reduce the inequality in earnings and investments, but by lowering supply elasticities widen the inequality in marginal rates, and thereby reduce efficiency.

Rising supply curves that are due to a segmented capital market create an inefficient allocation of the total investment in human capital by penalizing abler persons, who tend to invest more than others. If the market could not be made less segmented, efficiency could be increased by other, "second best," policies that favor the abler. The use of objective standards to select applicants for admission to subsidized investment institutions does favor abler persons, and could, therefore, offset the higher marginal cost of funds to them. Consequently, although objective selection may result in large inequality in earnings and investments, it could help allocate efficiently the total investment in human capital.[144]

Even severe critics of the distribution of incomes have generally protested only against unequal opportunities, and have treated in-

144 See section 9 of the mathematical appendix for a formal proof (p. 143).

equality resulting from differences in ability with indulgence, if not positive affirmation.[145] Possibly this simply reflects a basic philosophical distinction; I suspect, however, that partly reflected also is an implicit judgment about the interaction between equality and efficiency. By increasing the elasticities of and reducing the differences among supply curves, improved capital markets, scholarships to the needy, compensatory education, and the like reduce the inequality in earnings and at the same time generally improve the allocation of the total investment in human capital. The elasticities of and differences among demand curves, on the other hand, are less related to man-made factors, and more to the embodiment of capital in human beings, differences in intelligence, and other basic forces that are less easily corrected.[146]

5. *The Distribution of Property Income*

The discussion now shifts from the distribution of human capital and earnings to physical (that is, all nonhuman) capital and property income. Although earnings constitute 75 to 80 per cent of total incomes in the United States and are a major determinant of its distribution as well as level, property income should not be neglected in a serious study of income distribution. For one thing, most of the Anglo-Saxon literature on income distribution has stressed inheritance and property income even more than earnings.[147] Moreover, since property income is considerably more unequally distributed and skewed than earnings, its contribution to income distribution greatly exceeds that to income levels. Fortunately, the analysis developed in the previous sections seems capable of explaining why property incomes

[145] "Rightly interpreted, equality meant, not the absence of violent contrasts of income and condition, but equal opportunities of becoming unequal" (Tawney, *op. cit.*, p. 105); "But the greater the approach towards equality of opportunity, the more reasonable the contention that a distribution according to the value of work done is just" (Dalton, *op. cit.*, pp. 22–23); and in the meritocracy, "stratification has been in accord with a principle of merit, generally accepted at all levels of society" (Young, *op. cit.*, p. 99).

[146] Of course, even these forces could be offset by a system of taxes and subsidies, as, say, through a progressive tax on earnings. Since a progressive tax reduces inequality by redistributing earnings from the top earners, it discourages their investment in human capital more than others, and thereby tends to allocate the total investment in human capital less efficiently.

[147] For example, H. Dalton's classic study of inequality (*op. cit.*) devotes one chapter to incomes from work and seven to property income and inheritance.

are so unequal and skewed, and even why only a small fraction of the population appears to receive any inheritance.

The property income of any person can be divided into his "original" income and that due to his own capital accumulations. "Original" property income presumably comes from an inheritance and cannot be neglected as readily as "original" earnings were; it is discussed shortly. "Own" property income depends on the amounts invested and their rates of return, and assuming rational behavior once again, these two determinants in turn depend on the supply and demand curves for physical capital. The distributions of own property income and earnings differ, therefore, if and only if the distributions and elasticities of the supply and demand curves for physical and human capital differ.

There are no obvious reasons why the distribution of demand curves for physical and human capital differ significantly in any particular direction, but the other determinants are another story. Although the difference is sometimes exaggerated,[148] undoubtedly the market for physical capital is less segmented than is that for human capital. It was argued in section 4*d* that, while the net effect on inequality of greater segmentation is not clear a priori, it does cause lesser skewness.

Probably the major difference, however, is in the elasticities of demand. Section 2 argued that the marginal rate of return to an investor in human capital declines with increases in the amount he invests primarily because his human capital is embodied in himself and, therefore, his time is a crucial input into the investment process. Physical capital, on the other hand, is not embodied in people and generally the amount owned is not a major input into additional investments;[149] consequently, there is less reason to expect significant declines in the rate of return to any investor in physical capital as the amount invested increases.

Since demand curves for physical capital can be expected, therefore, to be more elastic than those for human capital, abler investors can be expected to specialize more in physical capital, that is, to invest considerably more than the less able investors.[150] As shown earlier, greater specialization causes the distributions of investments and in-

148 See my remarks on pp. 78–80.

149 The important exceptions are the examples beloved by the Austrians in their approach to capital theory: the aging of trees, wines, etc.

150 This partly explains why some accumulations of physical capital dwarf any accumulations of human capital.

comes to be more unequally distributed and skewed. Primarily, therefore, because of differences in demand elasticities,[151] although aided somewhat by differences in supply elasticities, our analysis implies that investments in physical capital and property income are more unequally distributed and skewed than are investments in human capital and earnings.

A person leaving an inheritance must decide how to distribute it between human and physical capital: a rational aim is to select that combination yielding maximum benefits. Since at least relatively small investments in human capital apparently generally yield considerably higher payoffs than those in physical capital,[152] one would expect the preponderant part of small inheritances to be placed in human capital, say in the form discussed earlier of gifts to children to finance their education, training, and health.[153] Since statistics and discussions of inheritance usually only include physical (that is, nonhuman) capital, probably most small inheritances are overlooked entirely!

As the amount inherited by any person increased significantly, his marginal rate of return on investments in human capital would decline significantly, and at some point would be brought into line with that obtainable on physical capital. The fraction going into the latter would then increase. As his inheritance increased still further, the marginal rates would decline on both forms of capital, but especially on human capital because its demand curves are less elastic. A larger and larger fraction would be placed in physical capital until, for extremely large inheritances, the preponderant part would be placed not in human but in physical capital.

To summarize, the analysis developed in this paper can predict a wide variety of facts about physical capital. Among them are why large accumulations of physical capital dwarf any of human capital,

151 I pointed out earlier that investments in human capital yield declining rates of return partly because the time remaining to collect returns becomes smaller and smaller (if lifetimes are finite). The result is that human capital is accumulated over a much shorter period of time than the length of a working lifetime. There is no such time limitation on physical capital, and it can be and is accumulated throughout a lifetime. Accordingly, the argument that property income is more unequally distributed and skewed than earnings because physical capital is accumulated for longer periods of time than human capital is partly a special case of our argument in terms of greater elasticities of demand.

152 See pp. 191–193.

153 Just as these inheritances were assumed to affect the supply curves of funds for own investment in human capital, so inheritances in physical capital could be assumed to affect the supply curves of funds for own investments in such capital. The supply of funds might then be more unequally distributed for physical than human capital because, as shown in the following, inheritances invested in the former are more unequally distributed than are those in the latter.

why own property income is more unequally distributed and skewed than earnings, why only a small and select part of the population appears to receive any inheritance, and why inheritances used for investments in human capital are less unequally distributed than those used for physical capital.

6. Summary and Conclusions

The main purpose of this essay has been to develop a theory of the distribution of earnings and to some extent of other income as well. Earnings are made dependent on the amounts invested in human capital, and the latter are assumed to be determined by a rational comparison of benefits and costs. In other words, each person is assumed to have a negatively inclined demand curve showing the marginal benefit and a positively inclined supply curve showing the marginal financing cost from an additional dollar of capital invested: the optimal capital investment occurs where these two functions intersect. The supply curve to an individual investor is positively inclined because the segmentation in the market for human capital forces him to tap more costly funds as he expands his capital investment. His demand curve is negatively inclined because the embodiment of human capital in investors makes their own time a major input into the investment process. The rise in the value of this time as capital is accumulated over calendar time, combined with the finiteness of working lives, eventually forces marginal benefits to decline as more capital is accumulated.

The distributions of earnings and investments would then depend on the distributions and shapes of these supply and demand curves. The "egalitarian" approach to income distribution, which neglects differences in demand and relates differences in supply to man-made differences in opportunities, implies that earnings are probably more equally distributed and less skewed than opportunities. The "elite" approach, on the other hand, neglects differences in supply and relates differences in demand to more intrinsic differences in capacities. It implies that earnings are more unequally distributed and skewed than capacities. A more general approach combines both differences in supply and demand—that is, in opportunities and capacities—and implies that earnings and investments are more unequally distributed and skewed the more elastic are the supply and demand curves, the more these curves are unequally distributed and skewed, and the greater the positive correlation between them.

The effects of various changes in opportunities on the distribution of earnings and on the efficiency of the allocation of the total investment in human capital were analyzed. Greater "equality of opportunity" not only tends to reduce the inequality in earnings, but also to increase the efficiency of the allocation. A less segmented capital market also tends to improve the allocation, but it increases the skewness and possibly also the inequality in the distribution of earnings. The use of "objective" criteria to select applicants for admisson to subsidized schools and other investment institutions could result in considerable inequality in earnings. "Objective" selection, by its favoring of abler investors, could, however, lead to a more efficient allocation since it could offset the penalty to abler investors from the positively sloped supply curves associated with a segmented capital market.

The analysis was also briefly applied to the distributions of physical (that is, all nonhuman) capital and the property income yielded. Since human capital is and physical capital is not embodied in investors, the demand curves for investment in the former tend to be less elastic than those for the latter. This can explain the observed greater inequality and skewness in the distributions of physical capital and property income than human capital and earnings. Inheritances appear to be received by only a small and select part of the population because small inheritances are invested in human capital and therefore are not reported in inheritance statistics. As the amount inherited by any person increased, a larger and larger fraction would be invested in physical capital. This can explain the sizable inequality in reported inheritances and can contribute to the large inequality in physical capital and property income.

The model developed in this essay can be looked upon as a special case of a more general model that includes uncertainty,[154] discrepancies between actual and desired capital stocks and investment rates, random shocks, and so forth. Although I believe that the special model is extremely useful in understanding actual income distributions, I have not tried to defend this view with any systematic empirical tests. Other empirical studies[155] do offer strong support for the relevance and significance of the theory developed in this essay.

[154] Reactions to uncertainty form the basis of the theory developed by M. Friedman in "Choice, Chance, and the Personal Distribution of Income," *Journal of Political Economy*, 71, 1953, pp. 277–290.

[155] See J. Mincer, *Schooling, Experience, and Earnings*, New York, NBER, 1974; B. R. Chiswick, *Income Inequality: Regional Analyses within a Human Capital Framework*, New York, NBER, 1974, and "The Average Level of Schooling and the Intra-Regional Inequality of Income: A Clarification," *American Economic Review*, 58, 3, June 1968, pp. 495–500; and Becker and Chiswick, "Education and the Distribution of Earnings," *American Economic Review*, 56, 2, May 1966, pp. 358–369.

An important attraction of this theory is that it relies fundamentally on maximizing behavior, the basic assumption of general economic theory. Moreover, at the same time, various "institutional" factors are incorporated: inheritance of property income, distribution of abilities, subsidies to education and other human capital, unequal opportunities, and other factors all have important parts in the discussion. The body of economic analysis rather desperately needs a reliable theory of the distribution of incomes. Whether or not this approach is ultimately judged to be satisfactory, it should demonstrate that such a theory need not be a patchwork of Pareto distributions, ability vectors, and ad hoc probability mechanisms, but can rely on the basic economic principles that have so often proven their worth elsewhere.

7. *Mathematical Appendix*

1. If the returns on each investment in human capital were constant for the whole remaining working lifetime, earnings of the j^{th} person at some age p after the investment period was completed would be

$$E_{jp} = X_{jp} + \int_0^{C_j} r_j(C) f_j \, dC, \tag{81}$$

where X_{jp} are the earnings at age p if there was no investment in human capital, $r_j(C)$ is the rate of return on the C^{th} dollar invested, f_j is a correction for the finiteness of working lives, and C_j is the total capital investment. If X_{jp} is small enough to be neglected, if working lives are long enough so that $f_j \cong 1$, and if the average rate of return is defined as

$$\bar{r} = \frac{1}{C} \int_0^C r(C) dC, \tag{82}$$

equation (81) can be approximated by

$$E_j = \bar{r}_j C_j, \quad \text{all } p. \tag{83}$$

The distribution of earnings would be related to the distributions of average rates of return and total capital investment, and to the interaction between them. Although some useful insights can be obtained from this relation,[156] the number is limited by the fact that average

[156] See pp. 83–88.

rates and capital investments are themselves both derived from more basic determinants of choice.

2. These determinants can easily be analyzed if investments are assumed to be carried to the point where the marginal rates of return just equal the marginal rates of cost. The function

$$\bar{r}_j = D_j(C_j) \tag{84}$$

gives the average rate of return to the j^{th} person, where

$$\frac{d\bar{r}_j}{dC_j} \leq 0. \tag{85}$$

By a well-known formula the marginal rate would then be

$$r_j \equiv \frac{d(\bar{r}_j C_j)}{dC_j} = \bar{r}_j\left(1 + \frac{1}{e_j}\right),$$

with

$$e_j \equiv \frac{dC_j}{d\bar{r}_j} \cdot \frac{\bar{r}_j}{C_j}. \tag{86}$$

Similarly the function

$$\bar{i} = S(C_j) \tag{87}$$

gives the average rate of repayment financing costs to the j^{th} person, where

$$\frac{d\bar{i}_j}{dC_j} \geq 0. \tag{88}$$

By the same formula the marginal rate of cost is

$$i_j \equiv \frac{d(\bar{i}_j C_j)}{dC_j} = \bar{i}_j\left(1 + \frac{1}{\epsilon_j}\right),$$

with

$$\epsilon_j \equiv \frac{dC_j}{d\bar{i}_j} \cdot \frac{\bar{i}_j}{C_j}. \tag{89}$$

Equilibrium requires equality between these marginal rates, or

$$i_j \equiv S_j\left(1 + \frac{1}{\epsilon_j}\right) = r_j \equiv D_j\left(1 + \frac{1}{e_j}\right), \quad \text{all } j. \tag{90}$$

Each of these j equations is a function of a single variable alone, C_j, and could, therefore, be solved for the set of optimal capital investments. Equation (84) would then give the set of optimal rates of return and (83) the set of earnings. Clearly the distribution of earnings would depend solely on the supply and demand curves, the D_j and S_j.

3. This dependency can be made explicit by assuming particular functional forms for these curves; a simple form that is also a first approximation to more general forms is the well-known constant elasticity function:

$$\begin{aligned} \bar{r}_j &= D_j(C_j) = a_j C_j^{-1/b_j} \\ \bar{i}_j &= S_j(C_j) = \frac{1}{\alpha_j} C_j^{1/\beta_j}, \end{aligned} \tag{91}$$

where α_j and a_j are constants and > 0, and b_j and β_j are constants defined by

$$\begin{aligned} b_j &= -e_j = \frac{-dC_j}{d\bar{r}_j} \cdot \frac{\bar{r}_j}{C_j} > 0 \\ \beta_j &= \epsilon_j = \frac{dC_j}{d\bar{i}_j} \cdot \frac{\bar{i}_j}{C_j} > 0. \end{aligned} \tag{92}$$

Equations (86) or (89) obviously imply that the marginal function has the same constant elasticity as the average one.

Substituting equation (91) into (90) gives

$$\frac{1}{\alpha_j}\left(1 + \frac{1}{\beta_j}\right) C_j^{1/\beta_j} = a_j \left(1 - \frac{1}{b_j}\right) C_j^{-1/b_j}, \tag{93}$$

or

$$C_j = \left(\frac{1 - \dfrac{1}{b_j}}{1 + \dfrac{1}{\beta_j}}\right)^{\frac{b_j\beta_j}{b_j + \beta_j}} (a_j\alpha_j)^{\frac{b_j\beta_j}{b_j + \beta_j}}, \quad \text{all } j. \tag{94}$$

Therefore, by equations (83) and (91)

$$E_j = \left(\frac{1 - \dfrac{1}{b_j}}{1 + \dfrac{1}{\beta_j}}\right)^{\left(\frac{b_j\beta_j}{b_j + \beta_j}\left(1 - \frac{1}{b_j}\right)\right)} \cdot (a_j\alpha_j)^{\left(\frac{b_j\beta_j}{b_j + \beta_j}\left(1 - \frac{1}{b_j}\right)\right)} \cdot a_j. \tag{95}$$

Note that for positive earnings $b_j > 1$, an important restriction on the elasticities that is used later on. The distribution of earnings would depend on the joint distribution of the four parameters b_j, β_j, a_j, and α_j.

To simplify the analysis still further assume that the elasticities of supply and demand are the same for everyone:

$$\begin{aligned} b_j &= b \\ \beta_j &= \beta. \end{aligned} \tag{96}$$

Then by using the notation

$$k(b, \beta) \equiv \left(\frac{1 - \frac{1}{b}}{1 + \frac{1}{\beta}} \right)^{\frac{\beta(b-1)}{b+\beta}}, \tag{97}$$

equation (95) can be written more simply as

$$E_j = k\alpha_j^{\frac{\beta(b-1)}{b+\beta}} a_j^{\frac{b(\beta+1)}{b+\beta}}. \tag{98}$$

4. Before discussing the general case given by equation (98), a few special cases are considered. If all supply curves were identical and had an infinite elasticity: $\alpha_j = \alpha$ and $\beta = \infty$, (98) becomes

$$E_j = k' a_j^b, \tag{99}$$

while (94) becomes

$$C_j = k^* a_j^b. \tag{100}$$

The distributions of earnings and investments differ only by a constant.

A log transform of (99) gives

$$\ln E_j = \ln k' + b \ln a_j, \tag{101}$$

and

$$\sigma(\ln E) = b\sigma(\ln a), \tag{102}$$

where σ is the standard deviation. Thus, the standard deviation of the log of earnings, a common measure of inequality, would be positively related to the elasticity of demand and to the standard deviation of the location of the demand curves. The shape of the distribution of E would also depend on b and on the distribution of a. For example, if a had a log normal distribution, so would E, and its skewness would be positively related to the size of b and the skewness of a. Indeed, since $b > 1$, both the skewness and inequality in E would exceed that in a.[157]

Moreover, the distribution of E could be positively skewed even if a was not, the more so the larger b. For example, if a was symmetric, the values of a both below and above its mean would be increased by the E transform (neglecting the new units k' and except, of course, for $a \leq 1$), but those above would be increased by greater absolute and percentage amounts. The result would be a stretching out of the larger values into an elongated tail, and the stretching would be greater, the larger b was.

To show this more explicitly, let $f(a)$ be the density distribution of a and $f'(E)$ be the corresponding distribution of E. Then by a well-known formula[158]

$$f'(E) = \frac{da}{dE} f(a) \tag{103}$$

$$= {}'k \frac{1}{b} E^{1/b-1} f(a). \tag{104}$$

If $\hat{a}$ and $\hat{E}$ are the modes of a and E respectively,

$$\hat{E} < k'\hat{a}^b, \tag{105}$$

which is evidence of the elongation of the right tail. The mode of E is found by differentiating equation (104) and setting it equal to zero. If a derivative is denoted by a dot $(\cdot)$ over the function differentiated, one has

$$\dot{f}'(E) = \left(\frac{'k}{b}\right)^2 (E^{1/b-1})^2 \dot{f}(a) + f(a) \frac{'k}{b} \left(\frac{1}{b} - 1\right) E^{1/b-2} = 0, \tag{106}$$

or

$$\hat{E}^{1/b} = \frac{b-1}{'k} \frac{\dot{f}(a')}{f(a)}, \tag{107}$$

where a' is the value of a that transforms into $\hat{E}$. Since $\hat{E}^{1/b}$, $'k$, $b-1$, and $f(a')$ must all be positive, so must $\dot{f}(a')$. But if a were a unimodal

[157] The simplest proof is to note that both the skewness and variance of a log normal distribution depend only on the variance of the normal distribution obtained by a log transformation (see J. Aitchison and J. A. C. Brown, *op. cit.*, pp. 8–9).

[158] See M. G. Kendall, *The Advanced Theory of Statistics*, London, 1945, Vol. I, pp. 16–18.

distribution, $\dot{f}(a')$ would be positive only if $a' < \hat{a}$, for $\dot{f}(\hat{a}) = 0$ by definition.

The discussion can be made more concrete by considering a couple of well-known distributions. If a was uniformly distributed, its density distribution would be

$$f(a) = \frac{1}{N_2 - N_1}, \tag{108}$$

and, therefore, from equation (104)

$$f(E) = \frac{'k}{b(N_2 - N_1)} E^{1/b-1}. \tag{109}$$

A uniform distribution is transformed into a monotonically declining distribution, the rate of decline being faster, the higher b. This long-tailed distribution has exactly the same shape as the Pareto distribution, except that the exponent in the latter is < -2, while $1/b - 1 > -1$; it is even closer to the distributions discussed by Zipf and Yule.[159]

If a was approximately normally distributed, the mode of E would be found from the relation[160]

$$a' = \frac{u + \sqrt{u^2 - 4(b-1)\sigma^2}}{2}, \tag{110}$$

where u is the mean and σ the standard deviation of a. If $b - 1 > \left(\frac{u}{2\sigma}\right)^2$, the mode of E would be at the origin and E would be another long-tailed distribution. For smaller b, E would rise to a peak[161] at a value for a' between $u/2$ and u, and then decline in a long tail.

159 For a comparison of several long-tailed distributions see H. Simon, "On a Class of Skew Distribution Functions," *Models of Man*, New York, 1957, Chapter 9.

160 A proof is based on noting that for a normal distribution

$$\frac{f(a)}{\dot{f}(a)} = \frac{-\sigma^2}{a - u}.$$

Then from equation (107)

$$a' = \frac{-\sigma^2(b-1)}{a' - u}.$$

161 Since the density of a is not zero when $a = 0$, the density of E would approach infinity as $E \to 0$ and would decline to a local minimum at a value of a equal to

$$a' = \frac{u - \sqrt{u^2 - 4(b-1)\sigma^2}}{2}.$$

5. If β was finite, equation (98) becomes

$$E_j = k'' a_j^{\frac{b(\beta+1)}{b+\beta}}, \tag{111}$$

while (94) becomes

$$\hat{C}_j = k^{*\prime} a^{\frac{b(\beta+1)}{b+\beta}} \tag{112}$$

The distributions of earnings and investments still differ only by a constant. Since

$$1 < \frac{b(\beta+1)}{b+\beta} < b, \tag{113}$$

the distribution of E and C would still be more unequal and skewed than a, but the differences would be smaller than when $\beta = \infty$. As $b \to \infty$, the exponent in (111) and (112) would approach $\beta + 1$.

6. If all demand curves were identical so that $a_j = a$, all j, earnings would be

$$E_j = k_i \alpha_j^{\frac{\beta(b-1)}{b+\beta}}, \tag{114}$$

while the amount invested would equal

$$C_j = k_* \alpha_j^{\frac{b\beta}{b+\beta}}. \tag{115}$$

These distributions are the same, aside from scale, only when $\beta = 0$ or $b = \infty$. Otherwise, since $b\beta > \beta(b-1)$, C would be more unequally distributed and skewed. Moreover, since $\beta(b-1) < b + \beta$ unless $b > 2$ and $\beta > b/b - 2$, E would very likely be less skewed and more equally distributed than α. A comparison of equations (111) and (112) with (114) and (115) shows that E and C would be more unequal and skewed for a given distribution of a_j, the demand curves, than for the same distribution of α_j, the supply curves.

7. If both a_j and α_j varied, E_j would be given by equation (98), and the variance of the log of E would be

$$\sigma^2(\log E) = \frac{\beta^2(b-1)^2}{(b+\beta)^2}\sigma^2(\log \alpha) + \frac{b^2(\beta+1)^2}{(b+\beta)^2}\sigma^2(\log a) + \frac{2b\beta(b-1)(\beta+1)}{(b+\beta)^2} R(\log \alpha, \log a)\sigma(\log \alpha)\sigma(\log a), \tag{116}$$

where R is the correlation coefficient between log α and log a. The inequality in E would be positively related not only to b and β and to the variation in a and α, but also to the correlation between the latter pair. The variance in log E would exceed that in either log α or log a unless $\sigma^2(\log \alpha)$ was much less than $\sigma^2(\log a)$, R was very negative, and b and β were rather small.

Note that the distribution of E (and of C) would be unaffected, aside from scale, by equal percentage changes in all a_j or α_j. Thus economy-wide changes in the cost of funds or the productivity of human capital that change all average rates of return or all average repayment costs by the same percentage could significantly affect average earnings and capital investments, but have no effect on the distributions around the averages. Therefore, the usual emphasis on skill differentials in discussions of the distribution of earnings is completely beside the point, in our model, if these differentials are measured by average rates of return.

The skewness in E would be greater, the greater the skewness in α and a, the larger b and β, and the larger R. For example, if a and α were log normally distributed, E would also be, with a skewness positively related to the variance of its log, which by equation (116) is positively related to R. Again, if log a and log α were perfectly positively correlated, they would be related by the constant elasticity formula

$$\alpha_j = g a_j^d, \quad g, d > 0, \tag{117}$$

and E_j could be written as

$$E_j = k a_j^s, \tag{118}$$

with

$$s = \frac{b(\beta + 1)}{b + \beta} + \frac{d\beta(b - 1)}{b + \beta}.$$

For a given distribution of a_j, equation (118) has the same shape as (111): if a_j was uniformly distributed, both are monotonically declining distributions of the Yule-Zipf class, while if a_j was normally distributed, both are skewed distributions with modes given by equation (110). The inequality and skewness in (118) however, always exceeds that in (111), the difference being greater the larger d, the elasticity of α with respect to a.

8. The contribution of investment in human capital to "profits" is

not measured by total returns alone, but by the difference between them and total repayment costs:

$$P_j = \bar{r}_j C_j - \bar{i}_j C_j. \tag{119}$$

The distributions of P and E are exactly the same, however, aside from scale, so that all the results in the previous sections apply to P as well as E. For a proof, simply first substitute the definitions in equation (91) into (119), and get

$$P_j = a_j C_j^{1-1/b} - \frac{1}{\alpha_j} C_j^{1+1/\beta}; \tag{120}$$

then substitute the optimal value of C_j, and have

$$P_j = n^{1 - \frac{1}{b}} \alpha_j^{\frac{\beta(b-1)}{b+\beta}} a_j^{1 + \frac{\beta(b-1)}{b+\beta}} - n^{\left(1 + \frac{1}{\beta}\right)} \alpha_j^{\left(-1 + \frac{b(\beta+1)}{b+\beta}\right)} a_j^{\frac{b(\beta+1)}{b+\beta}}, \tag{121}$$

or

$$P_j = n' \alpha_j^{\frac{\beta(b-1)}{b+\beta}} a_j^{\frac{b(\beta+1)}{b+\beta}} \tag{122}$$

Thus, aside from a difference in scale, P_j in equation (122) is exactly the same as E_j in equation (98).

9. To maximize total earnings from a given total capital investment, one

$$\text{Max } E = \Sigma E_j = \Sigma r_j C_j, \tag{123}$$

subject to $\Sigma C_j = C_o$

which gives as a necessary condition

$$\frac{\partial E_j}{\partial C_j} = \lambda, \quad \text{all } j, \tag{124}$$

where λ is the marginal rate of return; that is,

$$\frac{\partial E_j}{\partial C_j} \equiv r_j \equiv \bar{r}_j + C_j \frac{\partial \bar{r}_j}{\partial C_j} = \lambda, \quad \text{all } j. \tag{125}$$

Equation (125) can be expressed in terms of the underlying parameters as

$$'k \alpha_j^{\frac{-\beta}{b+\beta}} a_j^{\frac{b}{b+\beta}} = \lambda. \tag{126}$$

Equation (126) necessarily holds if either $\alpha_j = \alpha$, for all j and $\beta = \infty$, or $a_j = a$, for all j and $b = \infty$. If b and β are both positive and finite, and a_j and α_j both variable, then (126) can only hold if log a and log α are perfectly correlated, and related by the linear equation

$$\log \alpha_j = d + \frac{b}{\beta} \log a_j; \tag{127}$$

otherwise the equilibrium marginal rates of return, the left-hand side of (126), differ. One easily shows that the variance of the log of these marginal rates is smaller, the smaller the variances of log a and log α, and the larger the positive correlation between them.