Estimating the Returns to Schooling: Some Econometric Problems

Zvi Griliches

Econometrica, Volume 45, Issue 1 (Jan., 1977), 1-22.

ECONOMETRICA

VOLUME 45 January, 1977 NUMBER 1

ESTIMATING THE RETURNS TO SCHOOLING: SOME ECONOMETRIC PROBLEMS[1]

BY ZVI GRILICHES

This paper surveys various econometric issues that arise in estimating a relation between the logarithm of earnings, schooling, and other variables and focuses on the problem of "ability" as a left-out variable and the various solutions to it. It points out that in optimizing models the "ability bias" need not be positive and shows, using recent analyses of NLS data, that when schooling is treated symmetrically, allowing it too to be subject to errors of measurement and correlated to the disturbance in the earnings function, the usual conclusion of a significantly positive "ability bias" in the estimated schooling coefficients is not only not supported but possibly even reversed.

1. INTRODUCTION

MUCH OF RECENT APPLIED WORK in the economics of education has concentrated on estimating a version of the following equation:

$$y_i = \ln Y_i = \alpha + \beta S_i + X_i\delta + u_i \tag{1.1}$$

where Y is a measure of income, earnings, or wage rates, S is a measure of schooling, usually in units of years or grades completed, X is a set of other variables assumed to affect earnings, u is a disturbance, representing the other not explicitly measured forces affecting earnings, assumed to be distributed independently of the X's and possibly of S, and i is an index identifying a particular individual in the sample.

Having written down such an equation, which goes under the name of earnings or income-generating function, one can ask immediately a long list of questions about it, questions that can serve as an outline for a text in applied econometrics (though the order in which they are asked might be different):

(i) What is income?

(ii) What is schooling?

(iii) Why should the equation have this particular functional form?

(iv) What other variables should be included in the equation (among the X's)?

(v) Why should there be a relation like that in the first place? In other words: (a) What interpretation can be given to such an equation? (b) What interpretation can be given to the estimated β coefficient? (c) Can one expect it to be "stable" across different samples and time periods?

(vi) Given answers to the previous questions, how should it be best estimated?

(vii) Who cares?

[1] Presidential Address, Third World Congress of the Econometric Society, Toronto, August 22, 1975. I am indebted to the National Institute of Education and the National Science Foundation for financial support of the work on which this paper is based, to Bronwyn Hall and Stephen Messner for research assistance, to Gary Chamberlain, Richard Freeman, and Sherwin Rosen for many discussions on this and related topics, and to Arthur Goldberger, Jerry Hausman, Christopher Jencks, and Paul Taubman for comments on an earlier version. All remaining errors are my own.

Obviously, I don't intend to, nor am I able to, answer all these questions today. There is also no clear natural order among them. Instead, I'll indulge in a bit of autobiography (question (vii)), skip lightly over the first three questions and the one really hard one on this list (question (v)), concentrating on the more purely "econometric" questions of specification (question (iv)) and estimation (question (vi)), especially on what is to be done if one believes that an important variable has been left out from the equation, and present several empirical examples of such estimates, based on samples from the National Longitudinal Survey of young males in the United States.

2. BACKGROUND

My first published paper [**18**], building on the work of Theil [**46**], dealt with the consequences of leaving out a relevant variable from the equation one was estimating. Three years later [**19**], building on the work of Schultz and Becker, I produced one of the first estimates of the contribution of schooling to the growth of total factor productivity (in agriculture), using income-by-schooling weights to construct an appropriate quality-of-labor index. Almost immediately the possible issue of bias in using such weights arose to haunt these and subsequent estimates of sources of growth. While continuing to produce various estimates of the contribution of schooling to economic growth (cf. Jorgenson and Griliches [**34**]), I tried to validate them by including a schooling variable when estimating production function in agriculture [**20, 21**] and in manufacturing [**22**], and check them against possible biases by estimating earnings functions including various ability measures [**11, 23**, and **27**]. It is this last set of studies that I want to discuss with you today, using language and concepts that date back to my first paper on specification bias.

Before doing that, it is worth noting, however, that much of the work on earnings functions by others was rather differently motivated. A major line of work, going back to Miller [**36**], Houthakker [**30**], and Mincer [**37**], put the emphasis on estimating the average rate of return to individuals from additional schooling. Another, but related, line of work, going back at least to Friedman and Kuznets [**16**], culminating in Hanoch [**28**] and since then spreading into a vast river of econometric studies threatening to engulf us all, uses the framework of the earnings function to ask questions about discrimination, attributing "unexplained" group differences in earnings to discrimination or other named but not directly measured interferences with the market.

I am making these distinctions to point out that the issue of specification bias is much more serious for the last type of studies. In the context of my original work, left-out variables would not affect the estimate of the potential contribution of schooling to economic growth if (i) they were uncorrelated with schooling or if (ii) the relationship between them and schooling persisted into the forecast period. Even if (i) or (ii) did not hold exactly, the consequence of that for my estimate of β would be relatively minor. But studies which identify the "residual" with something particular such as discrimination, are much more dependent on the original equation having accounted for "*everything* else".

3. EARNINGS AND SCHOOLING

Before one can discuss the appropriate empirical counterparts to be used in estimating such an equation, one has to say something about its interpretation. It is simplest to think of it as a kind of "hedonic" equation for labor. At any one time there are a number of different types of labor in the market, differing in their productivity but otherwise reasonably easily substitutable for one another. Market transactions and competition among employers and employees will establish relative prices proportional to the different marginal productivities of different types of labor. There are well known difficulties with this view (cf. Lucas [**35**] and Rosen [**43**]), and I'll come back to discuss some of them below. But it is the only interpretation that makes a modicum of sense.

Having accepted such an interpretation, we shall want a measure of earnings that comes closest to some transaction price for a well defined quantity of labor and a measure of schooling that corresponds to the qualities that employers are willing to pay for. The first will lead us to use wage rates per hour or week, when available. Using annual earnings, or income per year, will confound market transactions with issues of labor-leisure choice and the more transitory effects of unemployment. The second desideratum, having an "output" measure of schooling, is almost impossible to attain and we shall have to settle for much less than that.

The actual situation on measures of earnings or wages is much more complicated than is outlined above. Each job and each person are multidimensional. Jobs differ in their fringe benefits, in their conditions of work, and in their opportunities for training and advancement. The on-the-job training issue has been the focus of Mincer's [**38** and **39**] work and I shall not pursue it further here, except to note that unless one has a whole life history of the wages of an individual, it is best to have data on people who have been full-time in the labor market for a decade or so, i.e., in Mincer's terms, one wants to have observations in the neighborhood of the "overtaking" point. Otherwise, one has to be rather careful in defining and measuring a relevant "experience" variable.

Since one views schooling and other forms of training as production processes for human capital, one would like to have independent output measures of such processes. But nobody believes that we can get close to it by having an elaborate examination and summarizing the results by one final grand test score. We are stuck therefore, with input measures of schooling, measures of the time spent in institutions that are called "educational". We should keep this discrepancy between desires and reality in mind when we come later on to interpret the results of our analyses.

The simplest model that summarizes such considerations can be written as follows:

$$Y = p_h H e^{u}, \tag{3.1}$$

$$H = e^{\beta S} \cdot e^{v}, \tag{3.2}$$

$$y = \ln Y = \ln p_h + \beta S + u + v, \tag{3.3}$$

where p_h is the market rental price of a unit of human capital which may vary over time and space, H is the implied unobserved quantity of human capital, while u stands for other, preferably random influences on wages. Equation (3.2) is an implicit production function for human capital with time spent in school (S) as the major input and other human capital augmenting influences such as differences in the quality of schooling, or differences in the efficiency (ability) with which the time in school was spent by different individuals, represented by the v variable. Most of the issues of "ability bias" and simultaneity can be discussed in terms of the content of the u and v variables and the relationship of S to them.[2]

Before we turn to that, however, note that the functional form of (3.2) implies increasing returns to S. As such, it only makes sense if the costs of S, both direct and in terms of income foregone, rise at least as fast or faster than the return from it. In fact, the functional form can be interpreted as arising from a cost function whose only components are the rising interest costs of foregone income (cf. Mincer [**39**, p. 19]). If foregone income is the only cost of an additional year of schooling, and if the increment in log income due to this additional year of schooling (β) were constant in perpetuity, then $\beta \approx (\partial Y/\partial S)/Y$ has the interpretation of the "rate of return to the investment in schooling". Whether such a formulation is adequate to the problem has been a source of much debate and of many attempts at improvement (cf. Ben-Porath [**5**], Rosen [**42**], and Wallace and Ihnen [**47**], among others). I'll come back to this issue briefly below, but for our purposes this simple interpretation will suffice for now. In any case, the empirical evidence for the semi-log form is reasonably strong (see Heckman and Polachek [**29**]).[3] We turn, therefore, to the consideration of the estimation problems raised by equation (1.1) or (3.3), specifically the content of the X's and the relationship of u to S.

4. ABILITY AS A LEFT-OUT VARIABLE

Let us assume, provisionally, that the true equation to be estimated is:

$$(4.1) \qquad y = \alpha + \beta S + \gamma A + u$$

where A is a measure of "ability" which we have ignored in our estimation procedure. Then, as is well known,

$$(4.2) \qquad Eb_{ys} = \beta + \gamma b_{AS} = \beta + \gamma \operatorname{cov}(AS)/\operatorname{var} S,$$

which leads to the conclusion that the simple least squares coefficient of ln Y on S is biased upward (relative to β). This is based on the following assumptions: (i) that "ability" has an independent positive effect on earnings ($\gamma > 0$) above and beyond its effect on the amount of schooling (correctly measured) accumulated; (ii) that the relationship in the sample between the excluded ability and included schooling variables is positive ($b_{AS} > 0$); and (iii) that this is the *only* variable that has been left out *and* that all the other usual least squares assumptions hold.

[2] I shall ignore the possible variations in p_h in what follows below.

[3] In our data there is almost no improvement in fit when the schooling variable is divided into six categories and separate dummy variables are substituted for the continuous measure.

The simplest way of dealing with this problem is to find a measure of "ability" and include it in such an equation. If one is willing to accept IQ, AFQT, or similar test score measures developed by psychologists as being relevant to our "ability" concept, one can (with quite a bit of effort) find samples of individuals for whom such test scores are available, estimate the correct equation, and infer the magnitude of the bias that would occur if the particular variable were left out from the equation. Much of recent work, to which I have also contributed [**11**, **23**, and **27**], has concentrated on producing estimates of the percentage "ability bias" in b_{ys}, based on estimates of $\gamma b_{AS}/\beta$.[4]

I have the feeling now that much of this work, including my own, was somewhat misdirected. There is no good reason to expect the "relative ability bias" to be constant across different samples or to generalize easily from one study to another and to the population at large. Even if γ is a constant, the magnitude of the relative bias will depend on β and b_{AS}. The estimated β will differ across studies, depending, for example, on the parameterization of the equation. If, as has become rather common following Mincer's work, β is estimated holding "experience" constant, it will tend to be higher than in studies that hold the age of the observed individuals constant. Since most measures of experience equal or are close to age minus schooling, their use will increase the schooling coefficient by an amount equal to the age coefficient, leaving everything else largely unchanged, including the estimated γb_{AS}. Thus, while the absolute bias may be the same, the relative bias will be smaller in "experience" based equations. Also, one may expect β itself to differ somewhat across age and other groupings. More importantly, the bias depends crucially on b_{AS}, the relationship of the left-out ability variable to the included schooling variable, which may differ greatly across time and across different samples. Selection rules as to who goes to college (for example) have changed significantly over time (see [**45**]). The relationship of ability to schooling within families (when we look into the behavior of brothers or twins) may differ significantly from that between families. And finally, different samples may be selected differently, cutting off significant portions of the variance of schooling, changing thereby b_{AS} and the implied estimate of the "ability bias."

An example of such dependence of the estimate of the "ability bias" on the parameterization of the equation is illustrated in Table I, which reports several estimates of the earnings function based on 1969 wage rates (per hour) of not-enrolled young men with valid IQ scores ($N = 1362$).[5] Two sets of estimates are presented, one using the age of the respondent (in 1969) as an independent variable and the other using a nonlinear measure of work experience, based on independent data on weeks worked since end of school or age 14, whichever is later. The latter measure ($XBT = \exp - 0.1 \times EXPERIENCE$ 69) is constructed along the lines suggested by Mincer [**39**] and its use improves significantly the fit of the equation without changing the estimated "ability bias" by much.[6] It is about

[4] This work has been recently surveyed by myself [**26**] and by Welch [**48**] among others.
[5] See the Appendix for more details on this data set.
[6] Using a cubic in experience gives about the same results.

.008 in the age-held-constant equation and about .006 in the experience-held-constant version. But the implied estimates of the *relative* (percentage) ability bias differ by a factor of four! The point here is not that the last estimate is necessarily the correct one, but rather that one's estimate of the "percentage bias" is model-dependent and hence not easily generalizable across different data sets and formulations.

TABLE I

ALTERNATIVE EARNINGS FUNCTION ESTIMATES
NOT-ENROLLED YOUNG MEN FROM THE NATIONAL LONGITUDINAL SURVEY
DEPENDENT VARIABLE: $LW69$[a] ($N = 1362$)

Equation Number	Coefficient of: (and Standard error) $S69$[b]	IQ	AGE	XBT[c]	R^2 SEE
1	.022		.048		.288
	(.005)		(.003)		.337
2	.014	.0023	.049		.294
	(.005)	(.0007)	(.003)		.336
4	.065			−.85	.309
	(.005)			(.05)	.332
5	.059	.0019		−.86	.313
	(.005)	(.0007)		(.05)	.331

[a] $LW69$ = log of wage rate per hour on current or last job in 1969.
[b] $S69$ = schooling completed, in years.
[c] $XBT = e^{-.1EXP69}$, $EXP69$ = work experience since left school or since age 14, in years. The effect of an additional year of experience is .062, at the mean of the sample, falling from .078 after one year of experience to .032 after 10 years of experience.

5. ABILITY AS AN UNOBSERVABLE

> "The difference of natural talents in different men is, in reality much less than we are aware of; and the very different genius which appears to distinguish men of different professions, when grown up to maturity, is not upon many occasions so much the cause, as the effect of the division of labor." Adam Smith, *The Wealth of Nations.*

Because of the variability of the various econometric results and the lack of agreement on the role of "ability" in such an equation and on how it should be measured, the debate has see-sawed between those (such as myself) who explain high estimates of "ability bias" as due to the interrelationship of "ability" with other left-out variables such as the quality of schooling, and those who explain low estimates of ability bias (such as Cardell and Hopkins [**8**]) as the result of the use of erroneous measures of ability. I shall come back to the first point of view below, after we consider the definition, measurement, and role of "ability" in somewhat greater detail.

Two polar views are possible. "Ability" is IQ, or something close to it, and the only problem is that our measures of it are subject to possibly large (test-retest) errors. If we had data on more than one test or on some other relevant instrumental variables, this would be a simple garden-variety "errors-in-the-variables" problem, to be solved by standard econometric techniques. The alternative view is that "ability", in the sense of being able to earn higher wages, other things equal, has little to do with IQ. It is "the thing with feathers" (cf. Emily Dickinson and Woody Allen [**1**]). It is an unobserved latent variable that both drives people to get relatively more schooling and earn more income, given schooling, and perhaps also enables and motivates people to score better on various tests. Basically, it is a hypothesis about the cause of and a re-interpretation of the correlation among the residuals from individual income, schooling, test scores, and other equations. As such, it is only loosely related to "ability" as it is commonly understood by psychologists.[7] It could just as well be "energy" or "motivation". To the extent, however, that test scores are admitted as "indicators" of such an unobservable, one can stake out some middle ground between these two extreme views.

Thinking this way leads one fast into simultaneous equations with errors of measurement models, a topic surveyed recently by Goldberger [**17**] and myself [**24**]. In such models the problem of identification is quite acute and requires a more careful and explicit specification of the way one assumes the data to have been generated. In particular, if one can assume that different observations, across equations or across individuals, share the same values of the unobservable variable, this will provide a source of "replication" and a way of identifying some of the parameters of the model. Such considerations explain the recent interest in data on siblings or twins (cf. Chamberlain and Griliches [**11**], Olneck [**40**], and Taubman [**44**]), since one would expect them to share some common components of the unobservable ability variable. Hence also the name of "variance-components" that keeps reappearing in such models.

At this point it is worthwhile to outline a more complete model of the whole process. It is not the most general model possible but it formalizes some of the arguments made in the literature and should suffice for the interpretation of the empirical examples to be discussed below.[8] We shall assume that:

(5.1) $\ln Y = y_1 = \alpha_1 + \beta_1 S + \gamma_1 A + u_1,$

(5.2) $LT = y_2 = \alpha_2 + \beta_2 SB + \gamma_2 A + u_2,$

(5.3) $S = y_3 = \alpha_3 + \beta_3 SB + \delta_3 B + \gamma_3 A + u_3,$

(5.4) $SB = y_4 = \alpha_4 + \delta_4 B + \gamma_4 A + u_4,$

(5.5) $T = y_5 = \alpha_5 + \gamma_5 A + u_5,$

(5.6) $A_{ij} = \delta_5 B + f_i + g_{ij}.$

[7] Becker's [**2**, p. 18] definition of "ability" as "earnings received when the investment in Human Capital is held constant" is close to this second polar view. Unfortunately, it is not operational since actual human capital accumulated and the ex-ante marginal rate of return to it are both unobservable. In our context it is easiest to identify Becker's view with the proposition that β differs across individuals.

[8] This model is an adaptation of the one outlined in Griliches and Mason [**27**]. See also Chamberlain [**10**].

This model contains alternative test and schooling equations, depending on the exact dating of the tests. Thus LT stands for "late test" which is itself a function of the schooling already achieved at the time of the test. Such a distinction between early (T) and late (LT) tests introduces also two schooling variables into the model and forces us to specify two separate schooling equations, one for current or expected schooling (S) and the other for early, before late test, schooling (SB). The unobservable ability variable (A) is interpreted as a measure of "initial human capital",[9] consisting of measured family effects ($\delta_5 B$), unmeasured family effects (f), and individual (g) random components.

In general, one would expect a positive correlation between B and f, but the model can be reparameterized so that f is defined net of B.[10] Note that the measured family background variables (B) are assumed to affect income only via schooling or the ability variables. This is an identifying restriction which may be testable if the model is overidentified enough.

I have said nothing yet about the source of the schooling equations ((5.3) and (5.4)). Together, they make completed schooling depend on social and regional background variables, the unobserved ability measure, and a random component. The important point here is that these equations formalize the argument for an "ability bias" in the least squares estimates of β_1 (given that γ_1, γ_3, and $\gamma_4 > 0$). Ideally, we would like to derive such equations from the optimizing behavior of individuals and I shall come back to this point, even if only briefly, below.

The estimation of such models depends crucially on what is assumed about the distribution of the left-out variable A and all the other disturbances (the u's). Assuming that A is the *only* left-out variable common to more than one equation and that the u's are distributed independently of A *and* of each other results in the most overidentified and simplest to interpret version of the model. In this model u_2 and u_5 are independent "nontransmitted" errors of measurement in the respective tests and the only source of bias in the least squares estimates of the schooling coefficient is due to the dependence of S on A. Some of these assumptions could be relaxed to allow for a dependence of S on u_5 (or a correlation between u_3 and u_5) which could be interpreted as "test-wiseness", something that makes one do better both on tests and in school but does not lead to additional income beyond its effects on schooling. It is possible, also, to test for the presence of more than one common factor in such models.[11]

If test scores are available, the simplest way of estimating equation (5.1) is to substitute one of them for the unobserved ability variable and use the background variables and the other test scores (if more than one is available) as instruments.

[9] This version can be reconciled with the one outlined in Section 2 by assuming that schooling augments an initial human capital measure: $H(t) = H(0)\, e^{\beta S(t)}\, e^{v(t)}$, that test scores are proportional to the logarithm of $H(t)$: $e^{T(t)} = CH(t)^{\gamma}\, e^{\varepsilon(t)}$, and by defining $A = \ln H(0)$. If ability were also to affect the rate at which human capital is accumulated per time unit spent in school, it would introduce additional interaction terms between A and S in the income equation.

[10] Given our data we cannot really get into the question of whether A is largely genetic or due to common environments. With data on twins or other relatives of different degrees, such a model could be extended to encompass this question also (cf. Behrman and Taubman [**4**]).

[11] Such extensions are explored in Chamberlain and Griliches [**12**].

Depending on which test score is available we get either

(5.1a) $y_1 = \alpha_1 + \beta_1 y_3 + (\gamma_1/\gamma_5) y_5 + u_1 - (\gamma_1/\gamma_5) u_5$ or

(5.1b) $y_1 = \alpha_1 + \beta_1 y_3 + (\gamma_1/\gamma_3) y_2 - (\gamma_1/\gamma_3)\beta_2 y_4 + u_1 - (\gamma_1/\gamma_3) u_3$

and use instrumental variables to estimate the coefficients of the endogeneous early (y_5) or late (y_2) tests variables. This is not fully efficient, but should give one consistent and reasonable estimates of the parameters of interest.[12]

An example of such an approach is given in Table II, which uses the logarithm of the median income of the expected occupation at age 30 (*EXLOMY*) as its dependent variable and expected schooling (*ES*) as the major independent variable. It is based on a sample of all young men with valid IQ scores, including those still enrolled in school in 1969 ($N = 3025$).[13] In this sample, and for these variables, including a measure of ability into the earnings equation reduces the estimated coefficient of schooling by about .003 to .009, depending on the measure of ability used. Allowing for measurement errors in the reported test scores increases their coefficients substantially, reduces the estimated schooling coefficients somewhat more, by another .007 or so, but doesn't change the major conclusions. The overall direct contribution of "ability" as measured by test scores to the explanation of the variance of individual expected earnings is quite small, on the order .01 (the adjustment for measurement error doesn't change this estimate; the increase in the estimated coefficient of ability is almost exactly offset by lowering the estimate of its "true" variance).[14]

The rather large swings exhibited by the "ability" coefficients when errors in measurement are allowed for imply that these measures are quite unreliable. Among the two tests scores available to us, IQ comes off a bit better, with an implied reliability of 0.7 as against only 0.5 for the KWW measure. That is, about 30 per cent of the observed variance of IQ is estimated to be in "error", at least as far as its impact on expected earnings is concerned. Given the unreliability of these measures it becomes important and interesting to impose more prior structure on our model and utilize the additional information that might be available in data on siblings. To the extent that siblings share some of the unobservables, they can effectively serve as "instruments" for each other.

The workings of a siblings model have been described in some detail in Chamberlain and Griliches [**11**] and Griliches [**24**]. I will not recapitulate it here except to note that the basic idea of that model is to take advantage of the somewhat different information contained in the between-families and the within-families variance-covariance matrices of all of the endogeneous variables.

[12] More efficient complete model methods are outlined in Chamberlain [**9**], Joreskog and Goldberger [**31**], and Joreskog and Van Thillo [**33**].

[13] The advantages and disadvantages of using such expected variables are discussed in the Appendix. The advantages briefly are: The data (i) are conceptually closer to the "overtaking" point; (ii) more relevant to models which postulate ex-ante optimizing; (iii) allow us to analyze a much larger sample, especially of brothers where the sample size is quite important to us.

[14] Note that the overall results are very similar to those in Table I using "actual" instead of "expected" data, holding experience constant.

TABLE II

ESTIMATES OF EARNINGS FUNCTION COEFFICIENTS FROM THE NLS YOUNG MEN SURVEY

DEPENDENT VARIABLE: *EXLOMY*[a] ($N = 3025$)

Estimation Method and Equation Number	*EXSC*[b]	*KWW*[d]	*IQ*[e]	*S66*[c]	R^2	*SEE*[i]
	Coefficient of[(f,j)]					
a. OLS	.068 (.003)				.182	.349
b. OLS	.065 (.003)	.0057 (.0010)		−.010 (.006)	.191	.348
c. OLS	.059 (.003)		.0028 (.0005)		.192	.347
d. TSLS:[g] *KWW* endog.	.057 (.004)	.0177 (.0022)		−.026 (.006)	n.a.[h]	.356
e. TSLS:[g] *IQ* endog.	.052 (.004)		.0051 (.0009)		n.a.[h]	.356

[a] *EXLOMY* = logarithm of median earnings of the expected occupation at age 30.
[b] *EXSC* = expected total schooling.
[c] *S66* = actual schooling in 1966.
[d] *KWW* = score on a test of the "knowledge of the world of work" (administered in 1966).
[e] *IQ* = score in high school.
[f] Other variables in the equation: age and dummy variables (*DATELOMY* and *DATE* 66) for dating the expectational variables.
[g] Additional instruments used in TSLS estimation: mother's education (*MED*), father's occupation when *R* 14 (*FOMY*14), home culture index (*CULTURE*), *SIBLINGS*, *RACE*, *REGION*, and *IQ* in the equation in which *KWW* appears and *KWW* in the equations when *IQ* appears.
[h] n.a. = not applicable.
[i] *SEE* = standard error of estimate (estimated standard deviation of the residuals).
[j] Numbers in parentheses are the estimated standard errors of the respective coefficients.

Table III presents maximum likelihood estimates of the sibling version of the model outlined in equations (5.1) and (5.6), assuming that the u's are all independent of each other.[15] "A" is interpreted to be a normal "ability" variable, affected by the two test measures and affecting the two schooling levels and possibly also the expected income measure. It is composed of a measured family

TABLE III

NLS BROTHERS: ESTIMATES OF EXPECTED EARNINGS AND SCHOOLING MODEL ($N = 580$)[a]

Equation and Dependent Variable	*EXSC*	*S66*	A	$\hat{\sigma}_u$
	Coefficients of			
1 $y_1 = EXLOMY$	.057		.034	.363
2 $y_2 = KWW$		.651	2.19	5.64
3 $y_3 = EXSC$		.368	.681	1.59
4 $y_4 = S66$			.322	.903
5 $y_5 = IQ$			6.88	9.10

[a] $\tau = \sigma_g^2/\sigma_f^2$, $\hat{\tau} = .72$, $\sigma_g^2/\sigma_A^2 = .21$.

[15] I am indebted to Gary Chamberlain for computing these estimates. See Chamberlain and Griliches [12] for more details and alternative versions. The computations were made using an amended version of the Joreskog, et al. [32] ACOVSM program. Since we partialed out first the age and "dates" variables, the resulting estimates are not fully efficient.

background component $B\delta_5$, an unmeasured family component f_i, and an individual independently distributed component $g_{ij}(A_{ij} = B\delta_5 + f_i + g_{ij})$.

All equations contain also the age, region of origin south, and dates variables. Background variables (father's occupation, mother's education, number of siblings, race, and "culture") enter in an unconstrained fashion in the schooling equation and are constrained to have proportionally the same coefficients in the ability variable with the constraint effective across the two test scores and the expected income equations. The simple least squares estimates of $b_{y_1y_3}$ holding age and dates constant is .075. It is .069 when estimated solely from the "within" brothers variance components (Σ). The comparable least squares estimates holding IQ constant ($b_{y_1y_3\cdot y_5}$) is .066, with the coefficient by IQ ($b_{y_1y_5\cdot y_3}$) estimated at .0024. The estimated asymptotic standard errors are about .008 for β_1 and .02 for γ_1.

On the whole, the results are very similar to the earlier ones (given in Table II). By allowing for errors in the test scores and by using the fact that we have two test scores *and* information on the family structure of the observations, we reduce the estimated coefficient of schooling by an additional .009 beyond the simple least squares estimates which already include an ability variable, while doubling the estimated *IQ* coefficient (from .0024 to .0049 (= .034/6.88)). The relatively small movement in the schooling coefficient is connected to the rather small direct role of the ability variable in the expected income equation. It accounts for only about three per cent of the estimated total disturbance variance in that equation. About 79 per cent of the total variance of the ability variable is due to common family components, both measured and unmeasured. This is higher than would be predicted by a purely-genetic-no-environment-effects model for brothers (.5 to .6). About 62 per cent of the estimated variance of these common family effects are associated with the measured background variables (father's occupation, mother's education, number of siblings, "culture" index, and race). The variance of the expected schooling variable can be apportioned as follows (after solving out the $S66$ equation): due to unmeasured "ability" components: 23 per cent (joint among brothers 13 per cent, individual 9 per cent); due to common measured family background (both direct and indirect via ability): 26 per cent; due to other individual effects uncorrelated with either "ability" or measured "background": 50 per cent. A similar decomposition can be computed for the IQ and KWW variables, indicating that the "pure" error components account for 34 and 62 per cent, respectively, of the observed variances of these variables.[16]

Since the estimated model is significantly over identified, one could relax some of the imposed independence assumptions and test several additional interesting hypotheses. But that would take us too far afield.[17] On the other hand, before we accept the above results as final there are at least two more arguments worth considering.

[16] All of these decompositions are net of the age and dates variables.

[17] See Chamberlain and Griliches [**12**] for such additional tests and an extension to the two-left-out-factors case. Adding another factor improves the fit significantly, reduces the implied error variance of IQ to 23 per cent and raises the schooling coefficient to .064.

6. ON OVER DOING IT: A DIGRESSION[18]

Most of the discussion above has been asymmetric. It has focused on thinking about potential *upward* biases in the estimated β and trying to guard against them by adding "ability" or other types of variables (such as "background") to the original earnings function. But excessive zeal can eaily result in serious downward biases in our estimated $\hat{\beta}$'s. This is particularly true if, as is most often the case, our measures of schooling (S) are far from perfect, and especially if they too are subject to random errors of measurement.

One doesn't usually think that such a well defined variable as "years of school completed" could be subject to much recording or recall error. But in cross-sectional household interview data all of the variables are subject to some error. Even if the errors are small, their effect will be magnified as more variables are added to the equation in an attempt to control for "other possible sources of bias". We may kill the patient in our attempts to cure what may have been a rather minor desease originally.[19]

A small but not unrealistic numerical example may be of help here. Let the true equation be

$$y = .1S^* + .01A + u$$

where S^* is "true" unobserved level of schooling. The observed level of schooling is $S = S^* + e$, with e random and independent of S^*, and $\lambda = \sigma_e^2/\sigma_S^2 = .1$. That is, we are assuming that ten per cent of the observed variance of schooling is due to random measurement error, a figure that is of the right order of magnitude (cf. Bishop [**6**]). For simplicity we shall assume that A is measured without error. Let $\sigma_S = 3$, $\sigma_A = 15$, and $r_{AS} = .5$. Then, the simple least squares coefficient of schooling, ignoring A, equals

$$\begin{aligned} Eb_{ys} &= \beta - \lambda\beta + \gamma b_{AS} \\ &= .1 - .1 \times .1 + .01 \times 2.5 = .115 \end{aligned}$$

while including A in the equation leads to

$$\begin{aligned} Eb_{yS\cdot A} &= \beta - \lambda\beta/(1 - r_{AS}^2) \\ &= .1 - .1 \times .1/.75 = .087, \end{aligned}$$

raising the downward bias due to errors of measurement by one-third. Now consider adding some more background variables (B) to the equation which really do not belong in it explicitly, affecting income only via schooling. Such variables, together with ability, may be correlated with schooling on the order of .7 or an

[18] This section can be thought of as a restatement of the point originally made by Friedman [**15**, pp. 85–89] in his "Digression on the Use of Partial Correlation Coefficients in Consumption Research" in the *Theory of the Consumption Function*. I am indebted to Gary Chamberlain for reminding me of this passage.

[19] This point has also been recently made by Bishop [**6**] and Welch [**48**]. It is worth restating, however.

$R^2_{S\cdot AB} = .5$. This would imply

$$Eb_{yS\cdot AB} = \beta - \lambda\beta/(1 - R^2_{S\cdot AB})$$
$$= .1 - .1 \times .1/.5 = .080,$$

or a doubling of the downward bias due to the originally rather small errors in the measurement of schooling. As more variables are added in, or as the range of schooling is restricted by considering only the within-family, between-brothers, or between-twins variance of schooling, the implied $R^2_{S\cdot AB\ldots x}$ rises towards $1-\lambda$ while the estimated schooling coefficient goes to zero. Clearly, the more variables we put into the equation which are related to the systematic components of schooling, and the better we "protect" ourselves against various possible biases, the worse we make the errors of measurement problem.

It is a sad fact that in doing empirical work we must continuously search for the passage between the Scylla of biased inferences due to left-out and confounded influences and the Charybdis of overzealously purging our data of most of their identifying variance, being left largely with noise and error in our hands. In a sense, we run into a kind of uncertainty principle: The amount of information contained in any one specific data set is finite and, therefore, as we keep asking finer and finer questions, our answers become more and more uncertain.[20]

7. THE ENDOGENEITY OF SCHOOLING[21]

There are (at least) two problems with simple least squares estimates of earnings functions. The first, and the one we have been discussing at length above, is the "ability" problem. That problem can be solved by either getting hold of a good ability measure or by estimating its effect in some errors-in-variables context. As serious, or perhaps even more so, is the second problem: Schooling is the result, at least in part, of optimizing behavior by individuals and their families. This behavior is based on some *anticipated* earnings function. To the extent that the "errors" (from the point of view of us as observers) in the ex-post and ex-ante earnings functions are correlated, they will be "transmitted" to the schooling equation and induce an additional correlation between schooling and these disturbances. This again suggests the use of simultaneous equations methods in estimating the coefficient of schooling in such equations.

Whether the second problem is really serious depends on one's view as to how close individuals are to guessing their own and the economy's future and to what extent their actions can be interpreted as optimizing. I would think that it is not as serious as it appears at first sight because of (i) the large influence of random, and probably unanticipated, events on the actual earnings experience of an individual and (ii) the large influence of parents, the state, teachers, and classmates on the actual level of schooling achieved by an individual, only part of which can be interpreted as the result of his own ex-ante optimizing behavior.

[20] For a vivid exposition of the uncertainty principle, see Bronowski [7, Chapter 11].

[21] This section parallels the discussion in Mincer [39, pp. 137–140] and the appendix to the second edition of Becker's [3] book. I am also indebted to Sherwin Rosen for many discussions of this topic.

It is hard to be more precise about this, beyond such brief generalities. To do more one has to set up an explicit model of individual optimal behavior *and* embed it in a *general equilibrium* solution for the economy as a whole. The first task has been the focus of much recent work in human capital theory initiated by Ben-Porath's [**5**] seminal paper. By and large, it hasn't led to any econometrically tractable results. To get simple closed-form solutions for the optimal schooling levels and on-the-job training trajectories requires extremely strong assumptions about functional forms and individual behavior.

I shall borrow from Rosen [**42**] a very simple model to illustrate the various problems that arise in this context. Assume that the individual is trying to maximize his wealth (W) at "birth" (i.e., before deciding on schooling):

$$W(S)=\int_s^{\infty} y(S, A, u)\, e^{-r(S+t)}\, dt \tag{7.1}$$

where $Y=y(S, A, u)$ is the anticipated earnings function which depends on schooling, ability, and other factors (u) unknown to us, but not to the chooser. The interest rate by which the individual discounts the future is r and t is the post-schooling time (experience) index. Note the strong assumptions: infinite life, no post-school investments or age effects, a constant rate of interest, earnings foregone as the only cost of schooling, and no subsidies or taxes. For this simplified model it can be shown (cf. Rosen [**42**]) that the marginal conditions are given by

$$\frac{\partial Y}{\partial S}\Big/ r = y(S\ldots) \tag{7.2}$$

which says that the present value of the marginal increment (in perpetuity) in income due to a change in schooling should equal foregone income per unit of time spent on schooling. The optimal individual level of S, say S^*, is the solution of this equation. If the y function has the form we have assumed earlier,

$$Y=e^{\beta S+\gamma A+u}, \tag{7.3}$$

then

$$\frac{\partial Y}{\partial S}=\beta y(S\ldots)=ry(S\ldots), \quad \text{and} \quad \beta=r, \tag{7.4}$$

which gives us the interpretation of β as "the" rate of return but unfortunately produces no S^*. This is the case where the "demand" and "supply" functions are perfectly parallel, and either S^* is at its upper limit or at zero, or it is undefined. For a nontrivial solution we need either diminishing returns in the human capital accumulation function or increasing borrowing costs.

A simple extension of the above model will get us part of the way. Assume that some of the costs of schooling are subsidized by somebody (parents or the state) by the amount TR. Then (7.4) becomes

$$\beta Y=r(Y-TR). \tag{7.5}$$

Substituting (7.3) for Y, taking logarithms, and solving for S gives

$$(7.6) \qquad S^* = \frac{1}{\beta}\left[-\log\frac{r-\beta}{r} + \log TR - \gamma A\right].$$

Note that in this formulation "ability" only affects the amount of initial human capital and, hence, the correlation between such ability and the optimal amount of schooling is *negative*! To get a positive effect of ability on schooling we have to allow it to interact with β, more able people accumulating more human capital per unit of schooling or, alternatively, let it lower the cost of schooling to the more able. The latter version would allow the more able to have more leisure or forego less income per unit of time in school or (more relevantly) per grade completed. If the real costs of schooling per unit of time are lower by the amount δA, then

$$(7.7) \qquad \beta Y = r(Y - TR - \delta A)$$

and

$$(7.8) \qquad S^* = \frac{1}{\beta}\left[-\log\frac{r-\beta}{r} + \log(TR + \delta A) - \gamma A\right]$$

with a possible positive net effect of A on S^*. Note that such an effect need not translate itself into a correlation with the disturbances in the earnings function. It makes schooling "cheaper" for the more able, but it doesn't lead employers to pay more for it given the achieved level of schooling.

This discussion also points out the possibility of a negative relationship between optimized schooling and the disturbance in the actual or expected income equation. Let us distinguish between expected income, foregone income, and "permanent" income, the piece that is joint to both concepts, and add another unmeasured individual income generating factor, μ_i, unrelated to the usual ability measure A (for example, "motivation" or "energy"). We can write "expected income" as

$$EY = Y_p \cdot e^u$$

where u's are the anticipated future differences in income that are unrelated to current income foregone, "net foregone income" as

$$FY = Y_p \cdot e^t - TR - \delta A$$

where t's are current transitory fluctuations which will not be transmitted into the future, and "permanent income" as

$$Y_p = e^{\beta S + \delta A + \mu}.$$

Then (7.8) can be rewritten as

$$(7.9) \qquad S^* = \frac{1}{\beta}\left[\log(TR + \delta A) - \gamma A - \mu_i - \log\frac{1}{r}(re^{t_i} - \beta e^{u_i})\right].$$

This messy expression is not intended to provide a realistic schooling decision estimating equation but only to illustrate the possiblity that schooling and the disturbance in the earnings function may be negatively correlated, leading to a *downward* bias in the usual least squares estimates of the schooling coefficient and implying that the schooling variable, too, should be treated as endogeneous and its coefficient estimated using simultaneous equations methods.

Such estimates are presented in Table IV for different subsamples of the NLS young men.[22] In addition to the ability variables, the schooling variables and associate variables such as experience are now considered endogeneous. The estimates which use the late test scores (*KWW*) in the equation use the early test score (*IQ*) and family background variables as instruments for all three right-hand side endogeneous variables (*EXSC*, *KWW*, and *S*66). The versions containing *IQ* use *KWW* as an instrument, instead. The latter is only legitimate if early schooling is assumed not to affect the later test. The results presented in Table IV, while not very precise, are quite surprising. They indicate that the original simple least squares estimates of the schooling coefficient may have seriously *under-estimated* rather than over-estimated it. Note also the drop in size and significance of the ability coefficients relative to the estimates presented in Table II. This is consistent with the view that they were "robbing" the schooling coefficient earlier. While the estimated coefficients of *S*66 are not very precise, they have the right sign in the

TABLE IV

EARNINGS FUNCTION COEFFICIENTS:
TSLS ESTIMATES WITH SCHOOLING ENDOGENEOUS
DIFFERENT NLS SAMPLES[a]

Coefficient of	Expected Variables Dependent Variable: *EXLOMY* $N=4601$	$N=3025$ a	$N=3025$ b	"Actual" Variables Not enrolled Dependent Variable: *LW*69 $N=1362$
EXSC (*S*69 for cols. 4 and 5)	.098 (.011)	.097 (.022)	.085 (.009)	.096 (.017)
KWW	.012 (.002)	.011 (.003)		
*S*66	−.071 (.014)	−.059 (.046)		
IQ			.0020 (.0011)	−.006 (.023)
SEE	.368	.356	.352	.339

[a] See notes to Tables I and II. Equations reported in Columns 1, 2, and 3 also contain age and dates variables. The equation in Column 4 contains, in addition, *XBT* (which is also treated as endogeneous), *RNS*, *SMSA*, and *BRNS*. Excluded instrumental variables in columns 2, 3, and 5 are the same as in Table II. Estimates in Column 1 use *also* a dummy variable for observations with missing *IQ* scores and interactions of all other instrumental variables with this variable.

[22] The brothers discussed in Table III are a subsample of the second set and, hence, are not discussed explicitly here. It is not obvious how to superimpose the endogeneity of schooling onto a family structure. A two-factor model would allow for a differential dependence between schooling and the disturbance in the income equation. Such a model yields .064 as its estimate of β_1. See Chamberlain and Griliches [12] for more details on this.

expected earnings and schooling equations, and indicate a positive effect of early schooling on late tests ($\beta_2 > 0$). In short, treating schooling and ability symmetrically turns most of our original conclusions around.

8. COMMENTS AND CONCLUSIONS

We have had only time to discuss a few problems raised by such estimates. We have, for example, devoted almost no attention to what is currently a major empirical gap in such models: the lack of explicit measures of on-the-job training and the use of "experience" as a kind of trend or "residual" measure to approximate them. We need a more direct attack on this problem and more specific measures of the phenomenon itself. Here and more generally, advantage should be taken of the time series nature of the various survey panels that have become available. All of the studies mentioned, including my own, have yet to approach such data from a truly dynamic point of view. Looking at the process of investment in human capital in greater detail should teach us more about the real world and help us to develop better "complete" models.

As we move in that direction, we shall soon wish to abandon the monolithic concept of schooling and of one human capital and start looking for different decisions and different pay-offs at different levels and for different types of schooling. Unfortunately, as we do that, our samples will start shrinking again and the quality of our inferences will deteriorate. Here too, models which postulate common unobservables appearing in different equations may help us to impose more structure on the data and reduce the information loss that occurs when we are tempted to slice our data sets thinner and thinner.

In addition, we still have a serious conceptual problem left. All such models take the $y(S, A \ldots)$ function as given. Since it reflects peoples' estimates of future market opportunities it should be similar across individuals, errors of forecast aside. To the extent, however, that the β's differ across individuals, we will have errors due to the mismeasurement of actual human capital accumulation by the observed years of schooling.[23] But the correct β at the macro-level will be still the slope of the current market opportunities locus, and not the ex-ante β's of particular individuals. Unfortunately, we have almost no workable theory of how such a locus is actually determined (cf. Rosen [**42** and **43**] for the difficulties we get into when we ask such questions). An interpretation that makes some sense argues that because different levels of human capital can be produced in the long-run with a relatively high supply elasticity according to the production function (7.3), even if the different types of human capital are not perfectly substitutable on the demand side, the resulting loci of ex-post intersections of demand and supply schedules will lie roughly along the long-run supply schedule. In the shorter-run, given the long leads and lags in the production and utilization of human skills, one should expect that particular groups will deviate from such a schedule according to

[23] If "ability" is interpreted as implying different ex-ante β's, then there should be an interaction between the schooling and ability measures in the earnings equation. We have not been able to detect such interactions in our data.

their shorter-run supply and demand elasticities.[24] In any case, there is no good reason to expect that such loci will remain constant over time. To go beyond such crude "hedonic" functions will require more explicit theories of the demand and supply structure for human skills.

In the mean time we may wish to summarize what we have learned from this particular excursion. Two theoretical points are worth reiterating: (i) In optimizing models there is no good a priori reason to expect the "ability bias" (or the direct coefficient of a measure of ability in the earnings function) to be positive. Thus, it shouldn't be too surprising if it turns out to be small or negative. (ii) An asymmetrical attempt to protect oneself against possible biases by putting more variables into the equation or by looking only within finer and finer data cuts, can make matters worse, by exacerbating other biases already present in the data. The empirical evidence examined here also points in the same direction: (i) Treating the problem asymmetrically and including direct measures of "ability" in the earnings function indicates a relatively small direct contribution of "ability" to the explanation of the observed dispersion in expected and actual earnings. The implied upward bias in the estimated schooling coefficient is about .01. (ii) Allowing for errors in measurement in such ability measures does little to change these conclusions except increase the estimated bias by another .005 or so. But (iii) when schooling is treated symmetrically with ability measures, allowing it, too, to be subject to errors of measurement and to be correlated to the disturbance in the earnings function, the conclusions are reversed. The implied net bias is either nil or negative. In addition, (iv) a more detailed examination of data on brothers indicates that if we identify "ability" with the thing that is measured (albeit imperfectly) by test scores, and if we accept the underlying genetic model which postulates that such a variable has a family components of variance structure, then the "unobservable" that fits these requirements seems to have little to do with earnings beyond its indirect effect via schooling.

In a sense, we have circled around our problem and data. We started looking for biases and at first found little. We kept on looking for more and leaned over more until we suddenly found ourseves on the other side of the original question. The whole process of such a research venture is perhaps best described by the following conversation between Pooh and Rabbit:

> "How would it be," said Pooh slowly, "if, as soon as we're out of sight of this Pit, we try to find it again?"
>
> "What's the good of that?" said Rabbit.
>
> "Well," said Pooh, "we keep looking for Home and not finding it, so I thought that if we looked for this Pit, we'd be sure not to find it, which would be a Good Thing, because then we might find something that we weren't looking for, which might be just what we were looking for, really."
>
> "I don't see much sense in that," said Rabbit.
>
> "No," said Pooh humbly, "there isn't. But there was going to be when I began it. It's just that something happened to it on the way."
>
> A. A. Milne, *The House at Pooh Corner*

Harvard University

[24] Cf. Freeman [**13** and **14**] for work along such lines.

APPENDIX

THE DATA BASE

The examples in the text are based on several subsamples from the National Longitudinal Survey of Young Men. In this survey, a random sample of about 5,000 young men was interviewed in 1966, followed up annually through 1971, and bi-annually thereafter.[25] At the moment only the 1966 through 1970 surveys are publicly available. Besides the usual economic and demographic variables, a test of the "knowledge of the world of work" was administered at the time of the initial interview (1966) and IQ-type test scores were collected from the high-schools of the respondents. I shall interpret the first test (*KWW*) as a test of "late" ability and the second (*IQ*) as a test of "early" ability. Unfortunately, the latter tests are unavailable for about a third of the sample, including all those who did not continue school beyond the ninth grade.

To save on sampling costs the Census Bureau based this sample and the parallel samples on the older men, mature women, and young women on one larger sample of households. Thus, there is some overlap of family members in the various surveys. In particular we have succeeded in identifying over 1,000 brothers, a subset of which is used in the analysis reported in the text.

From our point of view, the basic difficulty with this sample is the extreme youth of the members. As of 1969 close to half of the total sample was still in school. Moreover, those who were out of school and working were only about 22 years old, on average, and had only an average of four years of work experience (see Table V for details). Hence, it is hard to interpret their current status as being a good indication of their ultimate success in life. However, in addition to current status, the respondents were also asked about their expected total educational attainment and their expected ("desired" at age 30) occupation.[26] I have scaled (valued) their (three-digit) occupational expectations, by the median earnings of all United States males in 1959 in these occupations, converting it thereby into an "expected" income concept. Taking logarithms gives the major dependent variable (*EXLOMY*) used above.

The use of such "expected" variables has several advantages and disadvantages:

(i) It allows us to deal with expected income and schooling as of around the "overtaking" point and to ignore the difficulties created by the youthfulness of our cohort and the lack of explicit on-the-job training measures.

(ii) It comes close to dealing with the ex-ante optimizing behavior of individuals, as discussed in the Becker [**2**] or Rosen [**42**] models, uncontaminated by the ex-post encounter with reality.

(iii) Most importantly, it allows us to analyze almost the entire sample of individuals and to triple the available sample of brothers, avoiding also thereby the self-selection problem that would be posed by an analysis of only those who recently decided to stop their schooling.

The disadvantages are obvious:

(i) We are dealing with expectations and not "reality".

(ii) The use of occupational expectations as a proxy for income expectations ignores the expected returns to schooling and ability within occupations and the imposition of a uniform median income scale on the occupational expectations does not allow for differences in individual expectations about the differential future of various occupations.

(iii) The causality from schooling to earnings is much less clear for expectational variables.

Nevertheless, I believe that the expectational data are of intrinsic interest and that the advantages enumerated above outweigh the disadvantages. In any case, one gets rather similar results when "real" data for the "not-enrolled in 1969" subsample are used (see Tables I and IV).

[25] See Griliches [**26**] and Parnes, et al. [**41**] for more details on these data. They are based on a national sample of the civilian non-institutional population of males who were 14 to 24 years old in 1966. Blacks were over-sampled in a 3 to 1 ratio. The original sample consisted of 5,225 individuals, of whom 3,734 were white. By 1969 about 23 per cent of the original sample was lost, 13 per cent of it only temporarily (to the army).

[26] These are answers to questions "As things now stand, how much more education do you think you will actually get?" and "What kind of work would you like to be doing when you are 30 years old?" The first question was asked in every survey, the second only in 1966 and 1969. The latest available answers were taken and dummy variables were added for those observations which did not originate from the 1969 survey (*DATELOMY* and *DATE* 66, identified collectively as *DATES*).

TABLE V

DIFFERENT SUBSAMPLES OF YOUNG MEN FROM THE NATIONAL LONGITUDINAL SURVEY

MEANS AND STANDARD DEVIATIONS[a]

Variable	All	All Valid *IQ* Scores	Brothers (pairs) Total	Within	Not Enrolled in 1969 All	With Valid *IQ* Scores
N	4601	3025	580		2026	1362
AGE 69	21.2	21.5	20.3		22.2	22.3
	3.2	3.0	2.3	1.4	3.2	3.2
EXSC[b]	13.8	14.4	14.8		12.7	13.4
	3.0	2.4	2.3	1.1	2.8	2.3
S69[c]					11.6	12.5
					2.4	1.9
S66[d]	10.7	11.5	11.3		10.8	11.6
	2.4	1.9	1.7	1.1	2.4	2.0
EXLOMY[e]	8.61	8.65	8.67		8.53	8.58
	.403	.386	.404	.270	.389	.366
LW69[f]					5.60	5.68
					.426	.398
KWW[g]	33.3	35.5	34.9		33.0	35.1
	8.6	7.6	7.7	4.5	9.0	7.9
IQ[h]		101.2	102.8			97.7
		15.9	15.9	7.5		15.3
FOMY14[i]	5120	5372	5418		4826	5095
	1951	1960	2179		1779	1777
BLACK	.27	.17	.20		.28	.19
CULTURE[j]	2.2	2.4	2.5		2.0	2.3
	.97	.80	.76		1.0	.9
SIBLINGS	3.3	2.9	3.6		3.6	3.1
	2.6	2.3	2.1		2.7	2.4
EXP69[k]					4.0	3.7
					3.1	2.8
XBT[l]					.70	.72
					.27	.18
SMSA[m]			.67		.61	.65
ROS[n] or *RNS*[o]	.34	.33	.32		.41	.33

[a] The lower number in a pair of numbers is the standard deviation.
[b] *EXSC* = Expected total schooling to be completed eventually, in years.
[c] *S69* = Schooling completed in 1969, in years.
[d] *S66* = Schooling completed in 1966, in years.
[e] *EXLOMY* = Logarithm of the 1959 median earnings (in dollars) of all males in the occupation expected (desired) at age 30.
[f] *LW69* = Logarithm of hourly earnings (in cents) on the current or last job in 1969.
[g] *KWW* = Score on the "knowledge of the world of work" test, administered in 1966.
[h] *IQ* = Score on IQ-type tests, collected from the high school last attended by the respondent.
[i] *FOMY14* = Occupation of father or head of household when respondent was 14, scaled by the median earnings of all United States males in this occupation in 1959.
[j] *CULTURE* = Index based on the availability of newspapers, magazines, and library cards in the respondents home.
[k] *EXP69* = Post-school work experience. Estimated on the basis of the work record (in weeks) since 1966 and the date of first job after school and the date stopped school. Truncated at age 14, if respondent started working earlier. In years.
[l] $XBT = e^{-0.1 \times EXP69}$
[m] *SMSA* = Respondent in *SMSA* in 1969.
[n] *ROS* = Respondent in South when 14, columns 1–3.
[o] *RNS* = Respondent in South now (1969), columns 6–7.

We discuss the results for five subsamples: three subsamples based on expectational variables and two "not-enrolled in 1969" subsamples based on realized data on wage rates and schooling. The first sample is the largest; it contains almost all of the original sample ($N = 4601$). The main problem with this sample is that over 30 per cent of the respondents are missing IQ scores. The second sample is limited only to those having IQ scores ($N = 3025$). The third sample is a subsample of 290 pairs of brothers from sample two ($N = 580$). The fourth sample is a sample of all not enrolled young men in 1969 with valid wage and work experience data. It is a subsample of the first sample with $N = 2062$. The last sample is a subsample of the fourth sample, restricted to those respondents with valid IQ scores ($N = 1362$). Table V gives the means and standard deviations for the major variables in these samples and the definitions of these variables.

REFERENCES

[1] ALLEN, W.: *Without Feathers*. New York: Warner Books, 1975.

[2] BECKER, G. S.: "Human Capital and the Personal Distribution of Income," Woytinski Lecture No. 1. Ann Arbor: University of Michigan Press, 1967.

[3] ———: *Human Capital*. New York: NBER, 2nd edition, 1975.

[4] BEHRMAN, J. R., AND P. TAUBMAN: "Nature and Nature in the Determination of Earnings and Occupational Status," University of Pennsylvania, unpublished, 1975.

[5] BEN-PORATH, Y.: "The Production of Human Capital and the Life Cycle of Earnings," *Journal of Political Economy*, 75 (1967), 352–365.

[6] BISHOP, J. H.: "Biases in Measurement of the Productivity Benefits of Human Capital Investments," Institute for Research on Poverty, Discussion Paper No. 323-74, Madison, Wisconsin, 1974.

[7] BRONOWSKI, J.: *The Ascent of Man*. Boston: Little, Brown, 1974.

[8] CARDELL, N. S., AND M. HOPKINS: "The Influence of IQ on Income and on the Relationship Between Education and Income," Harvard University, unpublished, 1974.

[9] CHAMBERLAIN, G.: "Education, Income and Ability Revisited," Harvard Institute of Economics Research Discussion Paper No. 302, forthcoming in *Journal of Econometrics*.

[10] ———: "Unobservables in Econometric Models," unpublished Ph.D. dissertation, Harvard University, Cambridge, Ma., 1975.

[11] CHAMBERLAIN, G., AND Z. GRILICHES: "Unobservables with a Variance-Components Structure: Ability, Schooling and the Economic Success of Brothers," *International Economic Review*, 16 (1975), 422–449.

[12] ———: "More on Brothers," Harvard Institute of Economic Research Discussion Paper No. 469, Cambridge, Ma., 1976.

[13] FREEMAN, R. B.: *The Labor Market for College-Trained Manpower*. Cambridge, Ma.: Harvard University Press, 1971.

[14] ———: "The Declining Economic Value of Higher Education and the American Social System," Harvard Institute of Economic Research Discussion Paper No. 421, Cambridge, Ma., 1975.

[15] FRIEDMAN, M.: *A Theory of the Consumption Function*. Princeton, N.J.: NBER and Princeton University Press, 1975.

[16] FRIEDMAN, M., AND S. KUZNETS: *Income from Independent Professional Practice*. New York: NBER, 1945.

[17] GOLDBERGER, A. S.: "Unobservable Variables in Econometrics," in *Frontiers in Econometrics*, ed. by P. Zarembka. New York: Academic Press, 1974.

[18] GRILICHES, Z.: "Specification Bias in Estimates of Production Functions," *Journal of Farm Economics*, 39 (1957), 8–20.

[19] ———: "Measuring Inputs in Agriculture: A Critical Survey," *Journal of Farm Economics*, Proceedings issue, 42 (1960), 1411–1427.

[20] ———: "The Sources of Measured Productivity Growth: U.S. Agriculture, 1940–1960," *Journal of Political Economy*, 71 (1963), 331–346.

[21] ———: "Research Expenditures, Education, and the Aggregate Agricultural Production Function," *American Economic Review*, 54 (1964), 961–974.

[22] ———: "Production Functions in Manufacturing: Some Preliminary Results," in *The Theory and Empirical Analysis of Production*, ed. by M. Brown. Studies in Income and Wealth, Vol. 31. New York: Columbia for NBER, 1967.

[23] ———: "Notes on the Role of Education in Production Functions and Growth Accounting," in *Education and Income*, ed. by Lee Hansen. Studies in Income and Wealth, Vol. 35. New York: National Bureau of Economic Research, 1970.

[24] ———: "Errors in Variables and Other Unobservables," *Econometrica*, 42 (1974), 971–998.

[25] ———: "The Changing Economics of Education," Harvard Institute of Economic Research Discussion Paper No. 426, Cambridge, Ma., 1975.

[26] ———: "Wages of Very Young Men," *Journal of Political Economy*, 84 (1976), Part 2, 569–586.
[27] GRILICHES, Z., AND W. MASON: "Education, Income and Ability," *Journal of Political Economy*, 80 (1972), Part 2, 74–103; and in *Structural Equation Models in the Social Sciences*, ed. by O. D. Duncan and A. S. Goldberger. New York: Seminar Press, 1973.
[28] HANOCH G.: "An Economic Analysis of Earnings and Schooling," *Journal of Human Resources*, 2 (1967), 310–329.
[29] HECKMAN, J., AND S. POLACHEK: "Empirical Evidence on the Functional Form of the Earnings-Schooling Relationship," *Journal of the American Statistical Association*, 69 (1974), 350–354.
[30] HOUTHAKKER, H. S.: "Education and Income," *Review of Economics and Statistics*, 41 (1959), 24–28.
[31] JORESKOG, K. G., AND A. S. GOLDBERGER: "Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable," *Journal of the American Statistical Association*, 70 (1975), 631–639.
[32] JORESKOG, K. G., G. T. GRUVAEUS, AND M. V. THILLO: "ACOVSM: A General Computer Program for Analysis of Covariance Structures," Research Bulletin RB-71-1, Princeton Educational Testing Service, 1971.
[33] JORESKOG, K. G., AND M. V. THILLO: "Lisrel: A General Computer Program for Estimating a Linear Structural Equations System Involving Multiple Indicators of Unmeasured Variables," Research Bulletin RB-72-56, Princeton Educational Testing Service, 1972.
[34] JORGENSON, D., AND Z. GRILICHES: "The Explanation of Productivity Change," *Review of Economic Studies*, 34 (1967), 249–283.
[35] LUCAS, R. E. B.: "Hedonic Price Functions," *Economic Inquiry*, 13 (1975), 157–178.
[36] MILLER, H. P.: "Annual and Lifetime Income in Relation to Education," *American Economic Review*, 50 (1960), 962–986.
[37] MINCER, J.: "Investment in Human Capital and Personal Income Distribution," *Journal of Political Economy*, 66 (1958), 281–302.
[38] ———: "On-the-Job Training: Costs, Returns and Some Implications," *Journal of Political Economy*, 70 (1962), Part 2, 50–79.
[39] ———: *Schooling, Experience, and Earnings*. New York: NBER, 1974.
[40] OLNECK, M.: "Bias in the Income-Schooling Relationship: Some Results from the Kalamazoo Brothers Study," presented at the Third World Congress of the Econometric Society at Toronto. Madison, Wisconsin, unpublished, 1975.
[41] PARNES, H. S., et al.: *Career Thresholds*, Vols. 1–4. Columbus, Ohio: Center for Human Resource Research, Ohio State University, 1970–1973.
[42] ROSEN, S.: "Income Generating Functions and Capital Accumulation," Harvard Institute of Economic Research, Discussion Paper No. 306, Cambridge, Ma., 1973.
[43] ———: "Hedonic Prices and Implicit Markets," *Journal of Political Economy*, 82 (1974), 34–55.
[44] TAUBMAN, P.: "The Determinants of Earnings: Genetic, Family and other Environments: A Study of White Male Twins," University of Pennsylvania, mimeograph, 1975.
[45] TAUBMAN, P., AND T. WALES: *Mental Ability and Higher Educational Attainment in the Twentieth Century*. Berkeley, California: NBER and Carnegie Commission on Higher Education, 1972.
[46] THEIL, H.: "Specification Errors and the Estimation of Economic Relationships," *Revue de L'Institut International de Statistique*, 25 (1957), 41–51.
[47] WALLACE, T. D., AND L. A. IHNEN: "Full Time Schooling in Life-Cycle Models of Human Capital Accumulation," *Journal of Political Economy*, 83 (1975), 137–156.
[48] WELCH, F.: "Human Capital Theory: Education, Discrimination, and Life-Cycles," *American Economic Review*, 65 (1975), 63–73.