Sibling Models and Data in Economics: Beginnings of a Survey
Author(s): Zvi Griliches
Source: *Journal of Political Economy*, Vol. 87, No. 5, Part 2: Education and Income Distribution (Oct., 1979), pp. S37-S64
Published by: The University of Chicago Press
Stable URL: http://www.jstor.org/stable/1829908

Sibling Models and Data in Economics: Beginnings of a Survey

Zvi Griliches

Harvard University

This paper reviews a number of recent studies of the income-schooling-ability nexus using sibling data and discusses the problem of identification in such studies. Special emphasis is placed on the role of errors in variables, concluding that modest error levels can account for much of the observed difference between total and within-family estimates of returns to schooling. It also suggests that the family may not contribute as much to the transmission of inequality as is commonly thought, since it is a force for equality within (among siblings).

Science demands the submergence of social conscience in a welter of statistical squabbles.
[Harry Johnson, 1972]

I. Introduction

The rise of interest in the economics of education in the post–World War II period has also brought a renewed interest in the family as an economic decision unit and in data on siblings as a possible solution to the problem of unraveling the magnitude of "pure" economic returns to education from the various other forces which affect the observed

I am indebted to the National Science Foundation for financial support of this and related work, to Bronwyn Hall and Tom Stoker for research assistance, and to Michael Olneck and Paul Taubman for supplying me with unpublished data from their studies. I have also benefited from the comments of Yoram Ben-Porath, Gary Chamberlain, and Yoram Weiss. This paper was written during my stay at the Institute of Advanced Studies of the Hebrew University. It has been prepared for the Conference on Human Capital and Income Distribution at King's College, Cambridge, March 1978.

[*Journal of Political Economy*, 1979, vol. 87, no. 5, pt. 2]

individual income and schooling data. It is fair to say, I think, that most of the work in this area has used siblings data in an instrumental way, as an opportunity to "control" some variables that one could not do otherwise, and not in a substantive way. Little attention has been paid to the ways in which the family as an economic decision unit transmits both human and physical capital to its members and how this transmission mechanism affects the income distribution over the long run.

Brothers are likely to be more alike than a randomly selected pair of individuals on a variety of socioeconomic measurements.[1] This correlation arises from many sources: common heredity, both physical and cultural; similar access to financial resources; exposure to similar influences of friends, neighbors, schools, and other aspects of their particular community; the likelihood, even in adulthood, of closer location in space and hence exposure to similar regional price differentials and common business-cycle effects; and more. Some of these effects are measurable, but many are not, or only imperfectly so. This leads then to the expectation that in models of socioeconomic achievement the disturbances, which represent the force of all "other" unmeasured factors, will be correlated (positively) across brothers.

Differences between brothers are due to differences in their inherited abilities and motivations and to differences in their experiences and the stimuli they are exposed to. The extent to which environments differ between brothers depends on their age difference and on the changing circumstances of the family and community associated with this lapse of time. There are also interactions between differences in their inherited abilities and temperaments and the environments they chose or were exposed to.

An example of such an interaction would be the response of the family to perceived ability differences among its children. It can, by investing additional resources, either accentuate or attenuate them. If families want to reduce the inequality of outcomes among their children, they will try to compensate for the perceived inequalities in initial resource endowments by financing appropriate additional investments in human capital and/or changing their bequests appropriately. But such attempts are costly and will be limited by the level of resources controlled by the family.

Such considerations imply that differences in the schooling of brothers may be larger for brothers further apart in age and smaller for brothers from well-to-do families. There has been very little mod-

[1] Almost all of the work in this area has dealt with male siblings only, even though the models and econometric problems apply as well to sisters and mixed-sex pairs. While some work on sisters and brother-sister pairs is in progress, I shall follow the previous literature here and deal primarily with brothers.

eling of or evidence on this point, however. Instead, most of the recent work in this area has concentrated on the first aspect of siblings—their relative closeness in the space of unmeasured influences. A major focus of this work has been the attempt to eliminate potential biases in estimates of the effect of (returns to) schooling due to the presence of such unmeasured factors as "ability" or "family culture" by the use of differences between brothers as the basic source of information.

The earliest attempt to look at siblings data in economics known to me is Gorseline's (1932) dissertation, which set the pattern for most of the work to come by asking whether education paid when one contrasted the different educational experiences of brothers (156 pairs from Indiana, with data on 1927 income, schooling, occupation, and age). His results were used by Becker (1964), and the data were reanalyzed by Chamberlain and Griliches (1975). Various scattered pieces of data on siblings were synthesized and reviewed by Jencks et al. (1972). Recently there has been a rash of publications analyzing different sibling samples: Brittain (1977) analyzed about 60 pairs of Cleveland brothers; Chamberlain and Griliches (1977) analyzed 292 pairs of brothers from the National Longitudinal Survey (NLS) of Young Men with data on expected schooling and occupation and 161 pairs with actual (1970) wage data; and Jencks and his associates (see Corcoran, Jencks, and Olneck [1976] and Jencks and Brown [1977]) have analyzed data on 150 pairs of brothers collected by the National Opinion Research Center (NORC) in 1973 and on 99 pairs of brothers culled from the Talent Survey follow-up. Olneck (1976, 1977) collected and analyzed data on 346 pairs of brothers from Kalamazoo, Michigan, while Taubman and his associates (see Behrman and Taubman 1976; Taubman 1976; Behrman et al. 1977) have been analyzing a set of about 1,000 monozygotic (MZ) and 900 dizygotic (DZ) twin pairs based on the National Research Council (NRC) sample of white male army veterans.

It is not possible to summarize this rich set of detailed studies in a few words. Concentrating on what they have to say about the income-schooling relationship, one can divide them roughly into two groups: those who find only minor biases in the estimated returns to schooling due to omitted family background (Gorseline, Chamberlain and Griliches, and Jencks and some of his associates), and those who find that family background accounts for a major portion of the observed earnings-schooling relationship and is a major source of income inequality over time (Brittain, Olneck, and Taubman and his associates). This division is due in part to the way the question is phrased and in part to differences in methodology, but mostly to the fact that different samples appear to be telling different stories.

In what follows I shall first outline the econometric issues that arise when sibling data are used in an attempt to eliminate potential biases in the income-schooling relationship. The results of the major studies in this area will be reviewed next, and some additional calculations will be presented. Only at the end will I come back to consider some perhaps more interesting questions about the actual role of the family in all this, questions which have been largely ignored in this literature.

II. Simple Models

We start with a simple equation,

$$Y = \alpha + \beta E + \gamma A + u,$$

where Y is a measure of economic success such as the logarithm earnings, E is a measure of education, and A is an unmeasured left-out variable(s) such as ability, family status, or something similar. The problem of estimating β arises because of the lack of explicit or perfect measures of A and the probability that A and E may be correlated. Least-squares estimates of β which ignore A will be biased, picking up also some of the effects of A and attributing them to E. The standard left-out variable bias formula gives the size of this bias as $Eb_{YE} = \beta + \gamma\, b_{AE}$, where b_{AE} summarizes the relationship, in the sample, between the excluded A and the included E.

If the left-out variable is a purely "family" one, that is, if siblings have exactly the same levels of A, then it is clear that estimating β from within-family data, from differences between brothers in earnings and education, would eliminate this bias and give one the right β. This is the rationale, the promise, and the limitation of the siblings method. There are two major points to be made in this context:

1. It is unlikely that the A's we are interested in are purely "family" variables. Brothers' IQ scores, for example, are correlated, but only about .5. Taking differences will not eliminate the problem entirely. Similarly, if A is interpreted as a more general measure of initial (preschool) human capital, even MZ twins will not have exactly the same amounts of it. Once individual effects in A are allowed for, it is not clear anymore that within-family estimates are necessarily less biased. This will be shown below.

2. Leaving out A may not be the only thing that is wrong with the earnings function. There are, quite likely, several other serious problems with it, such as error of measurement in E (and other variables) and a possible simultaneity between the schooling decision and the earnings equation (correlation between u and E). It will be shown below that using within-family data may seriously aggravate such problems and cause opposite biases to increase.[2]

[2] See Griliches (1977) for a more detailed discussion of some of these issues.

This does not imply that sibling data are useless. On the contrary, they allow us to ask a richer and more complex set of questions. But one has to keep in mind that they are not a panacea and that simple within (between brothers or between twins) estimates are not necessarily closer to the "truth."

Let us now consider these points in turn. For this purpose I shall introduce notation which distinguishes between the family (i) and individual (j) components of the various variables. Letting

$$A_{ij} = f_i + g_{ij},\ H_{ij} = h_i + w_{ij},$$

$$E_{ij} = \eta\, A_{ij} + H_{ij},$$

$$Y_{ij} = \beta\, E_{ij} + \gamma\, A_{ij} + u = \beta\, H_{ij} + (\gamma + \beta\eta)A_{ij} + u_{ij},$$

we have a simple "complete" model with A being the variable that affects both E and Y directly, while H is a variable (or set of variables) which affects Y only indirectly via E. Both H and A have a family (f_i, h_i) and individual (g_{ij}, w_{ij}) components-of-variance structure. Also, H, A, and u are assumed, provisionally, to be mutually uncorrelated, and u is assumed not to contain a family component, the family structure being embodied solely in the two factors A and H.[3] As long as we have no explicit measures of A and H, we can define them to be uncorrelated with each other and without any loss of generality interpret H as that part of the other factors which is independent of A. Let us define the ratios (intraclass correlation coefficients) of family (between) components of variance to their respective total variances as ρ_A, ρ_H, and ρ_E, respectively, where $\rho_A = \sigma_f^2/\sigma_A^2$, $\rho_H = \sigma_h^2/\sigma_H^2$, and similarly, $\rho_E = (\eta^2\sigma_f^2 + \sigma_h^2)/(\eta^2\sigma_A^2 + \sigma_H^2)$.

Also, as long as we have not proposed yet a particular measure of A, let us set $\eta = 1$, and call the usual ordinary least-squares (OLS) estimator of β, b, and the within-families (differences between brothers) estimator b_w. Then

$$Eb = \beta + \gamma\, \sigma_A^2/\sigma_E^2 = \beta + \gamma/(1 + \mu),$$

$$\begin{aligned} Eb_w &= \beta + \gamma(1 - \rho_A)\sigma_A^2/(1 - \rho_E)\sigma_E^2 \\ &= \beta + \gamma(1 - \rho_A)/[(1 - \rho_A) + \mu(1 - \rho_H)], \end{aligned}$$

where μ is the ratio of σ_H^2/σ_A^2. A bit of additional algebraic manipulation makes it clear that the bias in b_w will be smaller than in b only if ρ_A

[3] Note that the education equation assumes that only total individual ability (A) matters, not its intrafamily distribution. A more general model, to be taken up in the last section of the paper, would rewrite it as $E_{iu} = \eta(f_i + \delta g_{ij}) + H_{ij}$; $\delta < 1$ would imply an attenuation of within-family ability differences as they work themselves through the schooling process. As long as $0 < \delta < 1$, the points made in this section remain largely unaffected by the generalization. Only if $\delta < 0$, i.e., if the family goes entirely counter to the innate differences, would one expect within-family estimates of the schooling coefficient to be more biased even in the absence of other determinants (H) of schooling.

$> \rho_H$, that is, if the family components account for a larger fraction of the variance in A than in H (or E). It is not obvious that this is necessarily true. Of course, it is true for the extreme case of only one left-out family factor, when $\rho_A = 1$ and $\rho_H = 0$. In the usual model where A is interpreted as IQ and H as a measure of other environmental influences on schooling, the purely genetic model implies a ρ_A of about .5 for brothers, but there is no reason to assume that ρ_H is necessarily less than .5. Hence, once individual effects are allowed for in A and family effects in H, that is, A is not just a family variable and not the only family variable, within-family estimates of β (b_w) are not necessarily less biased.

With these remarks as a background, we can now look at table 1, which summarizes the major results of recent sibling studies. Looking at columns 7 and 8, we see that with a few exceptions b_w is not strikingly different from b. Either the role of the left-out variable (γA) is not particularly large or its structure is not much more "familial" than that of the other variables in the model. There are some outliers, especially the Taubman MZ twins and the Cleveland brothers, but before we discuss them and other studies in greater detail, we have to outline the second major problem with these results: the possibility that using within-family data aggravates other biases in the data.

The problem can be put most simply as the possibility of a negative correlation between the schooling variable E and the disturbance in the earnings equations, u. In this case,

$$Eb = \beta + \gamma\, \sigma_A^2/\sigma_E^2 + \sigma_{uE}/\sigma_E^2,$$

and if the reasons for $\sigma_{uE} \neq 0$ are largely individual rather than familial, then going to within data will drastically reduce σ_E^2 without changing σ_{uE} much and make this source of bias much more important. If, as is likely, σ_{uE} is in fact negative, there will be a significant decline in b_w relative to b, which would be attributed to the importance of ability and "family background" but which in fact reflects nothing more than the aggravation of errors of variables and simultaneity problems associated with the schooling variable itself.

The point is easiest made in the context of the simplest random-errors-of-measurement model. Let measured schooling S be related to "true" education by $S = E + e$, where e is a purely random measurement error and hence without any familial content. Then

$$Eb_{YS} = \beta + \gamma\, \sigma_A^2/\sigma_S^2 - \beta\lambda_S,$$

where $\lambda_S = \sigma_e^2/\sigma_S^2$ is the ratio of error to total measured variance in schooling. The within-family estimate will be given by

$$EB_{(YS)_w} = \beta + \gamma\, \sigma_A^2\, (1 - \rho_A)/(1 - \rho_S)\, \sigma_S^2 - \beta\, \lambda_S/(1 - \rho_S),$$

TABLE 1

SELECTED SIBLING STUDIES: A SUMMARY

Survey, Year of Earnings Data, and Study	Sample Size (Pairs)	SD		Intraclass Correlation Coefficient			Estimated			
							b_{YS}		$b_{YS.IQ}$	
		ln Earnings	Schooling	ln Earnings	Schooling	Test Scores	Total	Within	Total	Within
Gorseline 1927 (Chamberlain and Griliches 1975)	156	.688	3.47	.37	.24	...	.082*	.080*	...	...
NLS 1969 (Chamberlain and Griliches 1977)	292	.404†	2.30†	.11	.51	.56	.074‡	.069‡	.066‡	.067‡
NLS 1973 (Griliches and Stoker)§	247	.441	2.44	.31	.55	...	.042	.039	...	...
Kalamazoo 1973 (Olneck 1976)‖	346	.446	2.73	.22	.55	.47	.067	.050	.049	.031
Talent 1971–72 (Corcoran et al. 1976)‖	99	.407	2.47	.21	.55	.58	.060	.057	.039	.042
NORC 1973 (Corcoran et al. 1976)‖	150	.870	3.11	.13	.53	...	.100	.110	...	...
NRC 1973, Twins (Behrman et al. 1977):										
MZ	1,022	.528	3.03	.54	.76	...	.077	.027	...	...
DZ	914	.567	3.13	.30	.54	...	.080	.059	...	...
Cleveland 1965–66 (Brittain 1977)#	52	.481	2.56	.40	.38	...	.078	.050	...	...

*Additional variables in the equations are age and age squared.
†Expected occupational income when 30 and expected schooling.
‡Additional variables are age, region of origin south, and dates (of the expected variables).
§ln wage rate; additional variables are age, SMSA, and region south; unpublished computations.
‖Coefficient estimates taken from Olneck (1977).
#Intraclass correlation coefficient from Brittain (1977, table 2-2). Estimated schooling coefficients from Brittain (1977, table 4-7). Earnings variance converted from log to ln and education variable divided by two to approximate years-of-schooling variable.

where ρ_A and ρ_S are the ratios of family to total variance components in A and S, respectively. It is easily seen then that while what happens to the "ability" bias is ambiguous, depending on the sign of $\rho_A - \rho_S$, the error-in-variables bias increases unequivocally by the ratio of $1/(1 - \rho_S)$. One can construct numerical examples (see Griliches 1977, p.12) which indicate that even for moderate λ_S this could be a serious source of bias. For example, $\lambda_S = .1$ (a value at the lower end of the range estimated by Bishop [1974] and by Bielby, Hauser, and Featherman [1977]), and the values of ρ_S given in table 1 imply a 20 percent downward bias in the within-brothers estimate of β and a 40 percent downward bias within MZ twins. The case of both S and IQ being measured with error will be taken up in the next section.

The algebra is essentially the same if the source of $\sigma_{uE} < 0$ is attributed to simultaneity and not just to errors in variables. Reasons why σ_{uE} may be negative are discussed at greater length in Griliches (1977). Here it will suffice to note that if u contains a permanent anticipated unmeasured human capital component and schooling is pursued optimally to add to one's stock of human capital, then one may expect a negative relationship between what one already has or anticipates to get in the future and the additional effort made to augment it further.

So far I have discussed this type of model and data assuming that one has no direct measures of "ability" and uses sibling or twin data to infer something about its impact (this is true of the Gorseline, NORC, NRC-twins, and Cleveland brothers data sets). But what if one does also have an IQ measure? If one is willing to identify A with IQ and there are no measurement errors in either IQ or E, then one can just estimate the $y = \alpha + \beta E + \gamma$ IQ equation and one does not need sibling data at all. Of course, these assumptions are not really tenable (see the discussion in Griliches [1977]), and hence while IQ data are valuable they are just not enough. Sibling data can help in identifying the structure of a much more complicated reality, but they too may not be able to give us the "truth" without requiring assumptions which at best are only approximately true.

III. Identification

Let us now turn from the discussion of possible biases to the question of the estimability of β (the returns-to-schooling parameter) in such contexts. Whether it can be estimated at all, and from what part of the data, will depend on what kind and how many variables are available in a particular data set and on what we are willing to assume about how they were generated. It is possible to distinguish several polar and intermediate cases:

1. No restriction is put on the type or number of family factors (unobservable variables) present in the model. Here estimation and identification must come entirely from the within-families (differences between brothers) variance-covariance matrices. This limits severely the scope of what can be done, since there are rarely enough exogenous (instrumental) variables left in the within-families part of the data, especially if one wants to allow for errors in schooling and IQ and/or simultaneity between schooling and earnings decisions. In the Taubman models, for example, the unobserved variables are either purely genetic or familial, and hence using differences between MZ twins will eliminate them entirely. This may still leave one, however, with the problem of measurement errors in the schooling variable. But in a simple model of sibling differences,

$$\begin{aligned} dy_2 &= dw + \epsilon, \\ dy_3 &= \beta dw + u, \end{aligned}$$

where dy_2 is the measured schooling difference; dw is the "true" schooling difference; dy_3 is the earnings difference; and ϵ, u, and dw are mutually uncorrelated, β is not identified without a further assumption about the $\sigma_\epsilon^2/\sigma_{dw}^2$ ratio. This type of model could be estimated if one had observations on two additional "success" variables, for example, dy_4 ("early occupation") and dy_5 ("late occupation"), and rewrote it as

$$\begin{aligned} dy_2 &= dw + \epsilon, \\ dy_3 &= \beta_3 dw + \delta_3 dy_5 + u_3, \\ dy_4 &= \beta_4 dw \qquad\qquad + u_4, \\ dy_5 &= \beta_5 dw + \delta_5 dy_4 + u_5. \end{aligned}$$

Now dy_2 can be substituted for dw in the dy_3 equation and dy_4 can be used as an instrument for it, yielding both an estimate of β_3 and of $\sigma_\epsilon^2/\sigma_{dw}^2$. Using this estimate of the error ratio one can get an estimate of β_4, β_5, and δ_5, yielding together an estimate of the "total" coefficient of schooling $\beta^* = \beta_3 + \delta_3(\beta_5 + \beta_4\delta_5)$. The β^* coefficient would become unidentified, however, if one were to allow for simultaneity (a correlation between dw and the u's), errors of measurement in the occupation variables, or the possibility that "ability" may also differ in MZ twins.

2. Pure error-in-variables models: This is the easiest case. If one has measures of both ability and schooling and the only problem is that they are subject to random measurement error (independent across siblings), one can use one brother's variables as instruments in the other brother's equations. This is equivalent to estimating the system from the family-components (cross-sibs) variance-covariance matrix. One can also allow for simultaneity if there are enough distinct mea-

sured family-background variables and one is willing to assume that they "work" only via the true IQ and true schooling variables.

3. An interesting special case, no IQ measure but errorless schooling, was described in Chamberlain and Griliches (1975). Their model can be written as

$$\begin{aligned} y_2 &= A + w, \\ y_3 &= \beta y_2 + \gamma_3 A + u_3, \\ y_4 &= \delta y_2 + \gamma_4 A + u_4, \end{aligned}$$

where y_4 is another measure of individual success (in their paper—occupation) and the u's and w are independent of each other and purely individual, all the family contribution being contained in the unobservable A and in whatever measured background variables there might be in the sample. This is a one-unobservable-factor model with no direct indicator of it, no IQ or other test-score measures. Instead, it adds a "consequences" equation, determining occupation or subsequent income. The idea for the identification of this model came from the observation that the biases in the schooling coefficient from leaving out "ability" should change consistently as one went from total to within-family estimates of the schooling coefficients in the two equations (y_3 and y_4), since the same variable is left out in both equations.

It is easier, however, to use Chamberlain's forthcoming (1979) "purged-IV" approach to indicate how one would identify such a system. First, use y_4 as a proxy for A in the y_2 equation, resulting in $y_2 = \pi_1 y_4 + \pi_2 w + \pi_3 u_4$, where $\pi_1 = 1/(\delta + \gamma_4)$, $\pi_2 = \gamma_4/(\delta + \gamma_4)$, and $\pi_3 = -\pi_1$. Since w and u_4 are purely individual, π can be estimated consistently by using the other brother's y_4 value as an instrument. Now form $\tilde{y}_2 = y_2 - \hat{\pi}_1 y_4 \simeq \pi_2 w + \pi_3 u_4$, and use it as an instrument for y_2 in the y_3 equation. Since u_4 is assumed to be uncorrelated with u_3, this yields a consistent estimate β in the context of a left-out unmeasured ability variable.

This procedure, as already pointed out by Chamberlain and Griliches (1975), is based on a variety of restrictive assumptions: only one unobservable with a family variance component, no errors in schooling, and no correlation between the individual disturbances. Some of these assumptions can be relaxed if the model is richer in variables and equations.

4. An almost "complete" model, encompassing the earlier models as special cases, consists of the following set of equations:

$$\begin{aligned} y_1 &= \mathrm{IQ} = & \gamma_1 A + \eta_1 H + u_1, \\ y_2 &= S = & \gamma_2 A + \eta_2 H + u_2 + e_2, \\ y_3 &= \ln y = \beta(y_2 - e_2) + & \gamma_3 A + \eta_3 H + u_3, \\ y_4 &= Z \quad = \delta(y_2 - e_2) + & \gamma_4 A + \eta_4 H + u_4, \end{aligned}$$

where A and H are two unobservable factors with the following variance-components structure:

$$A_{ij} = B_i\pi_1 + f_i + g_{ij},$$
$$H_{ij} = B_i\pi_2 + h_i + w_{ij},$$

where B is a set of observed family variables, such as father's occupation, mother's education, the number of siblings; $y_4(Z)$ is another indicator of "success," such as occupation or earnings in another year; and u_1 and e are pure errors in measurement, which are not "transmitted" to the y_3 and y_4 equations. Note that the notation has changed slightly from that used earlier. We may, provisionally, identify "A" as cognitive ability, the thing that is supposed to be measured by IQ; "H" may be thought of more in terms of motivational or environmental resource variables. It is also the carrier of potential simultaneity between schooling and earnings. This interpretation is only one out of a set of possible ones. It is shown in Chamberlain and Griliches (1977) that the coefficients of A and H are not separately identifiable in such models without the addition of a series of rather arbitrary normalizing restrictions such as the orthogonality of H from A (net of B), or the setting of $\eta_1 = 0$ and/or $\gamma_3 = \gamma_4 = 0$. But under rather broad circumstances β may be identified even if γ and η are not.

Several substantive issues arise in such models: (*a*) How many unobservable variables should one allow for? With four observed variables, the model outlined above allows for only two unobservable factors with family variance components. This is a testable hypothesis which is related to the rank of the between-families (cross-sib) variance-covariance matrix. We have found that two factors are adequate for the explanation of all the data sets we have examined to date.[4] (*b*) Can one use measured background variables as instruments in estimating the coefficients of schooling and IQ in the y_3 and y_4 equations? This is equivalent to assuming that $\pi_2 = 0$ and h_i is orthogonal to B_i (or that $\eta_3 = \eta_4 = 0$), that is, background variables work only via schooling and ability and do not enter into the earnings equation directly, and the second unmeasured family factor (h) is rather different from and independent of the kind of family influences which are measured by the usual type of background variables (father's occupation, mother's education, siblings, etc.). If

[4] Analyzing the Gorseline and NLS data, Chamberlain and Griliches (1975, 1977) found that two factors suffice, i.e., that the cross-sibs moments matrix can be approximated adequately by a rank-two matrix. We used canonical correlation (between all the endogenous variables and a set of family dummies) to test this hypothesis. Olneck's data yield similar results with an estimated squared third canonical correlation coefficient of .56, which is not significantly different from the .50 implied by the null hypothesis of no remaining family structure. In contrast, Taubman and his associates assume separate genetic factors for each equation (variable), which forces them into the use of within-MZ moments as their sole source of identification.

one is unwilling to grant such restrictions, one has to achieve identification entirely through restrictions imposed on endogenous variables. For this one needs at least two "consequences" variables. Even then the model as written is not identified. It turns out that we can have either errors in schooling or simultaneity but not both (at least with only two consequent variables).

The errors-in-schooling model is estimable if we assume that $\eta_3 = \eta_4 = 0$, that is, that the second factor does not enter the earnings equations directly. Then one simply substitutes measured IQ for A and uses brother's IQ and brother's schooling as instruments. Without errors in schooling ($\sigma_e^2 = 0$) we have a regular two-factor model which allows for some simultaneity between schooling and earnings via the presence of the second factor H. Such a model is described and estimated in Chamberlain and Griliches (1977). Its identification can be explicated using Chamberlain's (1979) previously mentioned "purged-IV" approach. One can use y_1 and y_4 in the y_2 equation as proxies for the two unobserved factors A and H, resulting in an equation of the form

$$y_2 = \pi_1 y_1 + \pi_4 y_4 + (1 - \pi_4 \delta) u_2 - \pi_1 u_1 - \pi_4 u_4,$$

which can be estimated consistently using the brother's y_1 and y_4 values as instruments (since the u's are assumed to be purely individual and hence uncorrelated across brothers). One can then form $\tilde{y}_2 = y_2 - \hat{\pi}_1 y_1 - \hat{\pi}_4 y_4 \simeq (1 - \pi_4 \delta) u_2 - \pi_1 u_1 - \pi_4 u_4$ and use it as an instrument for y_2 in the y_3 equation. The resulting instrumental variable ($\tilde{y}_2$) is correlated with y_2 (because it still contains u_2) but not with the left-out factors A and H (or u_3).[5]

If one wanted to allow for both errors in schooling and simultaneity, one would need to have at least three consequences variables (e.g., earnings in 3 different years). Having such data one would be tempted, however, to allow for some additional correlation between such variables, for example, for serial correlation in the transitory-earnings components, leading one back to an underidentified model.

In short, one cannot really hope to estimate very complicated models which allow for a variety of unmeasured influences. Either one has better data in which the errors-in-variables problem does not arise, or one has to limit oneself to a very short and specific list of left-out variables and ways in which they affect the model. Latent-variable models and estimation techniques are no substitute for good data and substantive restrictions.

[5] This is not how one would estimate it in practice. Full-information maximum-likelihood estimates of this model can be computed using the Joreskog LISREL program. The procedure outlined in the text is only intended to show where the identification comes from.

With these somewhat obscure remarks out of the way, we can turn toward an examination and reinterpretation of the results of the various studies.

IV. Review of the Evidence

In this section I shall briefly summarize the findings of the sibling studies listed in table 1. Chamberlain and Griliches (1975) reanalyzed the Gorseline (1932) data from the late 1920s on 156 pairs of brothers in the state of Indiana. There was very little difference between the total and within-families estimates of returns to schooling. Since the data base contained no direct measures of "ability" (e.g., IQ scores), an attempt was made to use reported occupation (scaled by the log average income in these occupations) as another indicator, and a maximum-likelihood procedure was developed to estimate the parameters of a model in which "ability" is an unobservable with both family and individual components of variance. The unobservable that emerged in that study appeared to be positively correlated across the income and occupation equations but negatively (though not significantly so) correlated with schooling. There is thus a commonality in the disturbances of the postschooling indicator variables, but it does not affect the estimated schooling coefficient in the income equation. Attempts were made to relax the no-correlation assumption between the disturbances in the income and occupation equations with no significant change in the results. Since tests indicated the presence of two rather than just one unobservable factors, a two-factor model (one factor purely familial) was estimated. This model is underidentified but yields bounds on β when other coefficients are constrained to lie within a reasonable range. Such bounds indicate that in these data there is little "ability" bias in the usual estimates of β. These findings are subject to at least three criticisms: (1) They are based on a rather old and unrepresentative set of data. (2) They use a rather dubious second indicator equation, occupational success, to identify the parameters of interest. It is not clear, however, whether one can really treat income and occupation as two different measures of success. And (3) since there are no test scores in this data set, the interpretation of the resulting unobservable variable as "ability" is rather tenuous.

Chamberlain and Griliches (1977) tried to respond to these criticisms by analyzing a more recent and more representative set of data on 292 pairs of brothers culled from the 1969 National Longitudinal Survey of Young Men, containing data on two types of test scores: IQ and a test of "Knowledge of the World of Work." Because of the youth of this cohort (17–27 in 1969), the study used expectational data on schooling and income (mean income of expected occupation at age 30). A two-unobservable-factors model similar to the one outlined in the previous section was estimated using both test scores as indicators of unmeasured ability. The results indicate little bias from the omission of such unobservable variables (maximum likelihood estimator [MLE] β = .064, versus an OLS β of .074 at the individual level). The implied unobservable factor which loads positively on the test scores and schooling appears to have no significant effect on expected occupational earnings net of expected schooling, while the unobservable that is correlated significantly with both expected schooling and earnings has opposite signs in the two equations. The latter left-out factor does not look like "ability" at all and implies a downward rather than an upward bias in the OLS estimates of β. These results can be criticized on at least three grounds: (1) Given the expectational nature of the data on schooling and income, it is not clear whether one can maintain the no-correlation-between-disturbances assumption even after the introduction of the two unobservable factors. A two-factor rational expectations type model implies the estimation of β from a smoothed between-families variance matrix. This yields a β of .061 with no significant change in the interpretation of the rest of the model. (2) The question may be raised whether actual data are anything like the expectations. To check on this, Chamberlain and Griliches estimated their model also for 161 pairs of out-of-school brothers with good wage, IQ, and work-experience data as of 1970. The full two-factor model did not converge, but a "one-and-a-half" factor model (second factor purely familial with no direct effect on income, $\eta_3 = 0$ and $\sigma_{u2} = 0$) again implied that the unobservable factor that is positively connected to test scores and schooling has no significant independent effect in the wage-rate equation (net of schooling and work experience). In this data set, too, there seems to be little bias from family-type effects, and whatever bias there may be does not seem to arise from the omission of IQ-type variables. (3) The youngness of this sample may be the reason why IQ-type variables seem to matter so little. There is some evidence that the importance of IQ in wage determination increases with age (see Hauser and Daymont 1977).

Griliches and Stoker have been following the NLS brothers as they age. As of 1973 there were 247 pairs of brothers out of school with good wage data. In this sample there was again no significant change in the estimate of the schooling coefficient when going from individual to within-families data.

Olneck (1977) collected and analyzed data on 346 pairs of brothers from Kalamazoo, Michigan, who were between 35 and 59 years old in 1973. Sixth-grade test scores were collected from school records, and data on current and past occupation, current earnings, and educational attainment were collected by phone interview. His estimate of the schooling coefficient is somewhat lower than in other studies, and its decline when estimated from the within data somewhat larger than usual but not out of line with the results of other studies. The big difference occurs in the importance that is assigned to the IQ variable. Once the within estimates are computed, including IQ, its coefficient goes up and the schooling coefficient drops significantly, to about less than half of its original value. Olneck concludes that the combined ability-background bias in the education-earnings relationship is quite large.

His results with respect to the IQ variable are not consistent, however, with the model outlined above or with most of the other models used in this area. A possible reinterpretation in terms of errors of variables will be presented below.

Jencks and his associates report (Corcoran et al. 1976; Jencks and Brown 1977) results for 99 pairs of brothers culled from the 11-year follow-up of the Talent Survey. The dependent variable is the logarithm of the wage rate at approximately age 28. The results are similar to those reported for NLS brothers—almost no decline in the schooling coefficient as one goes from total to within-family estimates, except that the estimated effect of IQ is higher. Holding IQ constant and looking at differences between brothers increases, strangely enough, the schooling coefficient—the opposite of what happens in the Kalamazoo sample. Since the sample is very small and heavily selected, it is probably not worth spending too much time on.

Jencks and his associates also report the results for 150 pairs of brothers collected by the NORC (earnings as of 1973). No test scores were available. The schooling coefficients are rather high and do not decline when estimated from differences between brothers. In summarizing their work on both samples (Talent and NORC), Jencks et al. conclude that the unobservable components that are common to the earnings of brothers have little to do with measured parental characteristics and are only weakly related to the unobservable family components in test scores and in schooling.

Taubman and his associates report (Behrman and Taubman 1976; Taubman 1976; and Behrman et al. 1977) on a large study of U.S. white male veteran twins, age 46–56 in 1973, based on 1,022 pairs of MZ and 914 pairs of DZ twins. The emphasis in their studies is on estimating "genetic" versus "environmental" sources of variation in education, occupation, and earnings. The model used is similar to that outlined in the previous section, except that A is interpreted as a purely genetic unobservable variable, while H is a purely familial environmental unobservable component. In their model the right estimate of β is to be had from the within-MZ-twins variance-covariance matrix, since both A and H are eliminated thereby. The resulting estimate of .027 is the lowest estimate in the whole table and also quite low relative to their estimate of the schooling coefficient at the individual level (.077 for MZ and .080 for DZ twins).[6] Their model also attributes about 45 percent of the observed variance in earnings to "genetics," 12 percent to other family-environment sources, and the rest (43 percent) to individual differences. There are several difficulties with the Taubman et al. studies:

1. It was shown by Chamberlain (1977*a*, 1977*b*) that data on twins do not provide any more identifying restrictions than data on brothers, unless one makes very special additional assumptions. A similar point is also made by Goldberger (1977). Behrman et al. assume that the nongenetic factors are purely familial and that the role of "environment" is the same for DZ and MZ twins. A basic asymmetry is postulated; MZ twins are assumed to be much "closer" to each other as far as genes are concerned, but this is not reflected in any greater environmental closeness or interaction. This does not seem to be a very attractive assumption, and there is some evidence against it (see Jencks and Brown [1977] and the last section of this paper). One would assume that parents and society treat children more alike the closer they are to each other in time, space, character, and appearance. The failure of this assumption removes the special identifying power attributed to the within-MZ data and leaves one with two samples with different ratios of σ_h^2/σ_H^2 with all the attendant identification problems.

2. As has been shown earlier (see also Bishop 1974; Griliches 1977), twin data, because of the high intraclass correlation in schooling, are especially susceptible to errors-of-measurement problems. A relatively modest adjustment for such errors can account for most of their results. I shall return to this point below.

3. As in most such studies, no allowance is made for the possible simultaneity between schooling and earnings.

4. The lack of test scores for most of their sample and the lack of a more explicit model of the sources of familial resemblance in the success patterns of twins and siblings make it difficult to interpret the latent variables that emerge and draw any clear policy conclusions from them. (This is not just a criticism of Behrman et al. It applies equally well to other studies which concentrate on the measurement of the "role" of family background.)

Brittain's (1977) study of 60 or so pairs of Cleveland, Ohio, brothers

[6] A more elaborate latent-variables model reported on by Behrman et al. (1977) comes down essentially to the same thing, with an estimate of β = .026.

is based on a sample of 659 decedent estates closed in 1964–65 and subsequent interviews with the heirs of these estates. The sample is the smallest and the most peculiar of all those reported on here. The correlation of income between brothers exceeds that reported for all other sibling sets except the Taubman-NRC MZ twins. Given the strange nature of this sample, I shall not comment on it further here.

There are several problems that plague sibling samples in general which should be mentioned here. (1) Most of the sibling samples are "opportunity" samples. The data had been collected for other purposes and are usually quite unrepresentative of the population at large (e.g., men in Indiana, Kalamazoo, or Cleveland; white army veterans; etc.). (2) "Brothers" as such may not be fully representative of the population at large. For example, they exclude only children and mixed-sex pairs. (3) There is also a serious sample-selection problem. Missing data on one brother tend to eliminate the whole pair from the sample.[7] To the extent that data are not missing entirely at random, it is quite likely that more of the discordant pairs are missing, overestimating the resemblance of brothers. Similarly, data are eliminated for nonresponse on the income or earnings question. To the extent that more low-income people are eliminated, this is likely to lead to an underestimate of the schooling coefficient. Methods are now available for tackling such problems explicitly (see Heckman 1976; Griliches, Hall, and Hausman 1978), but they have not been applied yet in this context.

V. The Effect of Errors in Variables

In this section I want to show that a reasonable estimate of the magnitude of errors of measurement in schooling can in fact account for most of the more extreme results observed in this literature.

[7] The seriousness of this problem is illustrated by the Taubman-NRC white male veteran twins sample, which was based originally on a census of all male twins born in the years 1917–27 in the United States (except for a few noncooperating states). The original population consisted of about 54,000 twin pairs. To appear in the current sample, both members of a pair had to have served in the armed forces, both had to survive to 1975, and both must have responded to all the questionnaires, both answering all the relevant questions on these questionnaires. The quantitative magnitude of the problem can be gauged from the following series of numbers: only 16,000 pairs (of the original 54,000) were located in army records. After several rounds of questionnaires, there were only 6,230 pairs with good addresses left by the mid-1970s. Though over a half of these responded to the recent survey, only in 39 percent of the pairs did both twins respond. Finally, only 1,900 pairs had good data on the major variables for both twins, leaving us with only a 3.5 percent yield from the original 54,000 population! While this is the largest siblings sample available currently, the heavy and repeated selection and self-selection throw serious doubt on the representativeness of the results derived from this sample.

In a recent study of replicated measurements, Bielby et al. (1977) give three estimates of the standard deviation of measurement error in census-type survey questions on years of schooling completed. They range between 0.6 and 1.8. If we take as our estimate of σ_e the midpoint of this range, that is, 1.2, and assume that this is the "inherent" measurement error in the type of survey instruments used to collect such data and that this is the only problem with such estimates, then we would predict the percentage decline in the within coefficient

$$\begin{aligned}(b - b_w)/b &= \rho_S \sigma_e^2/(1 - \rho_S)(\sigma_S^2 - \sigma_e^2) \\ &= 1.44\, \rho_S/(1 - \rho_S)(\sigma_S^2 - 1.44),\end{aligned}$$

which we can evaluate for different entries in table 1.[8] For Kalamazoo brothers the formula implies a decline of .29 versus .25 actual, while for the Taubman-NRC twins the predicted numbers are .59 for MZ and .20 for DZ versus .65 and .26 actual. They are of the right order of magnitude. Given what is known about the data-collection procedures of the Kalamazoo brothers and the NRC twin surveys, it is not unreasonable to assume that quality control was somewhat higher in the Kalamazoo survey (relative to average census practice) and probably a bit lower in the NRC survey, which would explain the relative under- and overshooting of such "predictions."

There are three problems with such back-of-the-envelope calculations: (1) In a number of surveys the change in the schooling coefficient from total to within is very small (NLS 1973, Talent, and NORC), while there is no reason to expect the schooling data to be that much more accurate there.[9] Something else must be going on. (2) The procedure does not take into account the effect of differencing between brothers on the magnitude of the ability bias. And (3) it does not allow for the possible effect of measurement errors in other variables such as IQ. Let us discuss the last two questions together. Consider a simple one-ability-factor model with IQ scores available but subject to error. At the individual level the true equation is

$$y = \alpha + \beta E + \gamma A + u;$$

E and A are unobserved. Instead we have the observed proxies $S = E + e$, $T = A + t$, where T is an IQ-type test score, and e and t are random and independent of each other and u. Then

$$y = \alpha + \beta S + \gamma T - \beta e - \gamma t + u,$$

[8] Note that we are assuming a constant absolute rate of measurement error.

[9] The predicted and actual are not too far off from each other for the Gorseline brothers, but for the NLS 1973, Talent, and NORC brothers the same procedure would predict a much larger decline than was observed. The NLS 1969 results are based on expectational data and hence do not really fit the model outlined above.

and the biases in the OLS estimates of these coefficients are given by[10]

$$\text{bias } \beta = E(b - \beta) = -\beta[\lambda_e - (\gamma/\beta)\, b_{TS}\, \lambda_t]/(1 - r^2),$$

$$\text{bias } \gamma = E(c - \gamma) = -\gamma[\lambda_t - (\beta/\gamma)\, b_{ST}\, \lambda_e]/(1 - r^2),$$

where λ_e and λ_t are the error ratios in S and T, respectively, b_{TS}, and b_{ST} and $r^2 = r^2_{ST} = b_{ST} \cdot b_{TS}$ summarize the relationship between the observed S and T in the sample. The question is what happens to these magnitudes as we move within families. Assuming that the relationship (η) between the two true magnitudes $E = \eta A + w$ is unaffected by it, we get for the within magnitudes (indicated by primes):

$$\lambda_e' = \lambda_e/(1 - \rho_S), \lambda_t' = \lambda_t/(1 - \rho_T), b_{ST}' = b_{ST}\,(1 - \rho_A)/(1 - \rho_T),$$

$$b_{TS}' = b_{TS}\,(1 - \rho_A)/(1 - \rho_S), \text{ and } (r^2)' = r^2\,(1 - \rho_A)/(1 - \rho_S),$$

where $\rho_A = \sigma_f^2/\sigma_A^2$. One can substitute these magnitudes in the above formulas, but they do not simplify much further. A plausible numerical example will illustrate instead the possible consequences and orders of magnitudes.

Assume that the original error rates in S and T are equal, $\lambda_e = \lambda_T = .1$. Assume that observed schooling is more "familial" than test scores: $\rho_S = .6, \rho_T = .4$, and hence $\rho_A = .4/.9 = .455$. Let us assume that $\gamma/\beta = .1$, which makes IQ slightly more important relative to schooling than is usually observed, and that $\sigma_T^2 = 255, \sigma_S^2 = 9$, and $r_{TS} = .5$ and hence $b_{TS} = 2.5, b_{ST} = .1$, and $\eta = .11$. With these numbers the relative biases in the estimates of β and γ at the individual level are $-.1$ and 0, respectively, while for differences between brothers these same numbers together with the assumed ρ values yield $-.29$ and $+.09$, respectively. That is, even though we start with the same rate of measurement error in both schooling and IQ, we do not get the same relative biases even at the individual level. Moreover, going to within data, the downward bias in the schooling coefficient is increased further while the bias in the IQ coefficient becomes positive. This is due to our assumption that the schooling variable is on net more familial than IQ ($\rho_S > \rho_T$), which is true in the Olneck data and may explain why the IQ coefficient increases within brothers. In the Talent data, the relationship between ρ_S and ρ_T is reversed and the IQ coefficient also falls. In any case, the fact that other variables are also subject to error (even at similar rates) does not imply that using within-families data will not aggravate this problem further.

[10] See Griliches and Ringstad (1971, Appendix C) for the derivation of similar formulas.

VI. Alternative Estimates: NRC Twins

In several of the data sets we can do better than just make back-of-the-envelope calculations with a priori assumptions about error variances. In the Taubman data set two additional indicator variables are available: indexes of early (first) and recent (1967) occupation. In the spirit of the latent-variable models estimated by Behrman et al. (1977), we can assume that there is one purely genetic latent variable A and another purely familial environmental variable H. The schooling variable is subject to error. The occupation variables are not. The earnings function is part of a recursive system with schooling affecting early occupation, schooling and early occupation affecting later occupation, and schooling and later occupation affecting earnings. The identifying restriction here is the exclusion of early occupation from the earnings equation in the presence of late occupation. By hypothesis, going within MZ twins eliminates both unobservable factors, the purely genetic A and the purely familial H. The only remaining problem is errors in S. It can be solved, within the framework of this model, by using early occupation as an instrument for schooling.

I present first the OLS estimates for the Taubman within-MZ twins data:

$$y = \underset{(.0078)}{.0174\,S} + \underset{(.0070)}{.0307\,O_2},$$

where the numbers in parentheses are the estimated standard errors of the respective coefficients. The instrumental variable estimates (using O_1 as an instrument for S) are

$$y = \underset{(.0576)}{.0511S} + \underset{(.0157)}{.0224\,O_2}.$$

To get at the reduced-form schooling coefficient, we need also to estimate the two occupation equations. But first we have to retrieve the estimate of the implied error variance in schooling, since without it the occupation equations are not identified. The error variance is computed by assuming that the IV estimate is "true" and using the relationship

$$b_{\text{OLS}} = b_{\text{IV}}\,[1 - \sigma_e^2/\sigma_S^2\,(1 - r_{SO_2}^2)],$$

or

$$.0174 = .0511\,[1 - \sigma_e^2/2.154\,(1 - .0765)].$$

This yields $\hat{\sigma}_e^2 = 1.312$ and $\hat{\sigma}_e = 1.145$, surprisingly close to the $\sigma_e = 1.2$ assumed in our earlier example. Subtracting 1.312 from the observed variance of schooling (within MZ twins) we can now compute:

$$O_2 = .741\,(S - e) + .094\,O_1,$$
$$O_1 = .535\,(S - e).$$

Together these yield the final estimate of the schooling coefficient

$$\hat{\beta} = .0511 + .0224\,(.7413 + .0943 + .5346) = .069.$$

This estimate is quite imprecise; the correlation between the instrument O_1 and S is only about .2, but it is now close to the range of other estimates. It allows for genetic variables, left-out family-environmental variables, and errors of measurement in schooling. The estimated error variance is 1.3, well within the range of the Bielby et al. estimates, implying that 61 percent of the observed variance of schooling within MZ twins is due to error of measurement. The same error accounts for only 14 percent of the variance among MZ individuals. Adjusting for error at the individual level would have yielded a schooling coefficient of .09. This coefficient is "free" of the errors-in-variables bias but is subject to the omitted-variables (A and H) bias. The latter is then estimated at about .02, only half of which is observed because of the already superimposed measurement-error bias.

These numbers should not be taken too seriously. First, they are based on a very weak instrument and are extremely imprecise. And second, they make rather strong assumptions, such as no measurement error in the occupation variables. Unfortunately, a more general model is not identified within the confines of their data set. Nevertheless, these computations show that their initial extreme results can be brought into line with other estimates by allowing both reasonable estimates of error of measurement in schooling and also letting the data "speak for themselves." There may be an "ability" bias in the usual OLS estimates of the schooling coefficient, but it is probably much smaller than is implied by their simple within-MZ regression estimates.[11]

VII. Alternative Estimates: Kalamazoo Brothers

The Kalamazoo (Olneck 1977) data set contains early IQ scores and hence allows the estimation of somewhat more elaborate models.

[11] In their chap. 7, Behrman et al. (1977) present the results for a somewhat more general latent-variables model with measurement error in schooling. The main difference is that their procedure utilizes also the DZ data at the cost of making additional assumptions. They estimate the reduced-form $\hat{\beta}$ = .078, which implies a somewhat higher $\hat{\sigma}_e^2$ = 1.5 instead of our estimate of 1.3. They dismiss these estimates as unreasonable, even though $\hat{\sigma}_e$ is within the range of the Bielby et al. (1977) estimates.

Treating IQ as a measure of the major "ability" variable, we are left with four possible problems: measurement errors in IQ, measurement errors in schooling, possible simultaneity between schooling and anticipated earnings, and the influence of additional unmeasured family factors. If the problem were just errors in measurement, we could estimate the coefficients from the cross-sib matrix. This yields an estimate of .098 for the schooling coefficient and a negative coefficient for IQ. The IQ variable has surprisingly little explanatory power at the between-family level. A more general one-ability-factor model with errors in schooling estimates the earnings equation at the individual level using brother's IQ, early occupation (in the presence of late occupation), and family-background variables as additional instruments and yields .063 for the coefficient of schooling holding current occupation constant (and still a negative IQ coefficient) and .084 for the reduced-form (total) coefficient of schooling. These estimates, however, are not responsive to the major claim of brothers and twins model: the elimination of all sources of bias from family-background-related variables. Unfortunately, if one allows for simultaneity at the individual (within-family) level, the schooling coefficient is not identified. There are not enough exogenous variables or restrictions to estimate it solely from the within matrix, allowing for both errors in IQ and a possible correlation of schooling with the earnings disturbance. One has either to impose outside estimates of error components or have samples with many more and different indicators of individual success.

VIII. Provisional Summary

The literature surveyed can be characterized as ranging between two extreme positions, one saying that family effects are various and many and that they lead to serious overestimation of the true return to schooling which should be estimated solely from differences between siblings, preferably MZ twins. The other extreme would take the position that family effects work almost entirely via schooling and hence cause little bias in the estimate of its coefficient, and that the decline observed in within-sibling estimates is the result of the aggravation of other problems, such as errors in measurement and simultaneity and not the reflection of true family effects. The evidence at this point is ambiguous. Some sibling studies show little change in the schooling coefficient, some show much. The latter can be explained by reasonable rates of measurement error, but this then leaves the former unexplained. Unfortunately, data are not currently available which would allow the estimation of a metamodel nesting both of these positions within it.

The following appears to be a fair summary of the state of our knowledge on this topic: Measured parental characteristics (except for race and region) appear to affect earnings primarily via their effect on the level of achieved schooling. The market does not appear to pay for them directly.[12] Unmeasured "family" characteristics have a substantial though not large effect on the variance of earnings (10–15 percent), but their interpretation is obscure. In particular, they cannot be interpreted as reflecting "visible" parental status effects such as wealth, which were already proxied to some extent by the measured demographic variables. They consist of all the influences on which brothers are "closer" to each other than a randomly drawn pair from the same population. But such influences include race, region, spatial proximity, similarity in the school system and peer groups of origin, similarity in the cultural environment which includes influences which extend far beyond the original "family" boundaries, and similarity in subsequent environments. They should not be interpreted, except tautologically, as reflecting the force of family influences. They reflect the force of accidents of birth into a family, race, region, city, ethnic group, and more, and the fact that spatial and social mobility within half a generation is not perfect. Even with much mobility, people on average move only some distance from their original location in geographic and social space. The correlation between brothers' earnings reflects this and should not be interpreted as being solely the effect of "family," or "class," nor as implying that there is no escape from one's "background." Note that all the background (and "foreground") measures taken together do not explain much more than about 30 percent of the observed variance of log earnings. If one were to make a major adjustment for the effect of transitory earnings, one would still be left with the conclusion that the inequality (opportunity?) within families is about as large as that between families. Whether the glass is half full or half empty depends on the point of view of the observer.

Note that up to now I have not talked much about R^2s or the "importance" of this or that variable. First, R^2s or "importance" depends on a comparison base, and that changes from sample to sample. Second, from a policy point of view, what matters are the slope coefficients, the derivatives with respect to particular variables of interest, and not "the fraction of variance accounted for." Nor have I

[12] This is the conclusion of the Corcoran et al. (1976) paper. In our own NLS studies we can accept the restriction that measured background variables work primarily via schooling and the "ability" factor. See Chamberlain and Griliches (1977, p. 107).

tried to break down the family effect into "genetics" versus "environment," since even if this were possible, it is not clear that it would give us any more handles on the problem without a fuller understanding of the mechanism by which these forces exert their influence. Some genetic differences might be corrigible (e.g., by enzymes), while for some environmental influences, such as cultural background, there are no effective intervention methods known.

Whether unmeasured common-to-brothers effects bias significantly the schooling coefficient is not clear. The direct evidence is ambiguous. If some weight is given to the possibility of errors of measurement and endogeneity of schooling, then much of the asserted bias disappears.

What good, then, are sibling data if by solving one problem (omitted familial variables) they aggravate other problems? While no panacea, they do provide us with a richer set of observational data and allow us to complicate our models and ask for more detailed answers. In econometric terms, they give us a larger set of instrumental variables, but only at the cost of explicit restrictions on the structure of our models. Something more is learned, but as is often the case in science, much of what is learned extends the dimensions of the unknown.

IX. What Do Families Do?

Almost all of the discussion above was in terms of what light sibling data can throw on estimates of returns to schooling. But there are other, perhaps more interesting questions that can be asked of sibling data. How does being brought up in a family of particular size and type affect the subsequent success of individuals? What is the role of the family in transmitting inequality from generation to generation? Does it accentuate or attenuate the inequality in native ability?

I think that such questions are very interesting, and that very little substantive work has been done on them by economists, myself included. Since my own reading in this area is quite thin, the comments I shall make may be quite superficial and are only intended to serve as an introduction to topics to be pursued further on another occasion.

The effect of family size and sibling position has been a matter of some interest to sociologists and psychologists. Recently there have been several works which outline theories of the negative effect of family size on the intellectual (human capital) growth of individuals as a consequence of the dilution of the family's economic and intellectual resources. Zajonc (1976) presents a theory of the development of individual IQ as a function of the average IQ (mental age) of the child's environment. Lindert (1974) discusses similar ideas in terms of

TABLE 2

ESTIMATED EFFECT OF A 1-SD DIFFERENCE IN IQ ON COMPLETED SCHOOLING (in Years), DIFFERENT SIBLING SAMPLES

DATA SET	NO. OF PAIRS	INDIVIDUALS: Total	INDIVIDUALS: Holding Measured Background Variables Constant	WITHIN: Sibling Differences
Siblings:				
Kalamazoo brothers	346	1.58	1.24	.90
NLS (1970) brothers*	292	1.11	.97	.74
Talent:†				
Brothers	205	1.06	...	.35
Sisters	205	1.08	...	.41
MF	457	1.03	...	.43
Twins:				
Talent:				
DZ, MF	93	1.25	...	.41
DZ, M	26	1.15	...	.07
DZ, F	51	1.12	...	.03
MZ, M	76	1.29	...	.13
MZ, F	90	1.09	...	.02
NRC:‡				
DZ	78	1.53	...	.23
MZ	196	1.59	...	.05

NOTE.—M = male, F = female, MF = mixed-sex pair.
*Expected schooling.
†From Jencks and Brown (1977).
‡GCT rather than IQ, Navy subsample.

the allocation of family resources (mainly time) to individual children as a function of their age and number. Both theories draw the implication that middle children in large families will suffer the most, and that the "sibling deficit" depends not only on numbers but also on their spacing. The data examined seem to support these theories, but the observed effects are rather small (relative to other sources of variance in educational and economic achievement).

A more interesting question, at least to me, is whether parents try consciously to attenuate the innate inequality between their children. I believe that families in fact act as (potential) income equalizers.[13] They try both to allocate resources equally between their children and to compensate, to some extent, for the handicaps of the children with lower natural endowments.

The available data sets, summarized in table 2, support this notion by indicating a significant lower effect of IQ on schooling within

[13] At least in recent generations. The historical role of primogeniture and the economic forces that led to changes in inheritance rules are worth exploring further.

families. Families attenuate the effects of native inequality in ability.[14] These efforts appear to be stronger (more successful) the closer the brothers are to each other in time and social space. Note in table 2 the much lower within coefficient of IQ for twins than for brothers in general.[15] Note also that mixed-sex twin pairs are treated almost like regular siblings, while same-sex twins have much lower response coefficients, even though their genetic variance is about the same. In the Talent data, there is no significant difference between the response coefficients of MZ and same-sex DZ twins, indicating that families treat close and same-sex siblings similarly. In the NRC data the coefficient for MZ twins is lower than for DZ twins, again indicating an interaction between the "environmental" response to IQ differences and the closeness of siblings in time and in resemblance. Such attempts to reduce within-family inequality between "close" children are likely to be costly, and only higher-income families may be able to afford them. Thus, I would expect that the within-family variance in socioeconomic achievement would decline at higher income levels.

Determining the net effect of all this on inequality over time requires the elaboration of a general equilibrium intergenerational wealth and human capital transmission model à la Conlisk (1974) which would incorporate in it the conscious efforts of parents to affect the within-family inequality of their offspring. That is a large task for the future. In the meantime, I would like to hazard the guess that Knight (as quoted by Johnson [1972]) may have been wrong in identifying the family as the major and persistent source of inequality. At

[14] This appears to go counter to hypotheses and interpretations of Becker and Tomes (1976). They interpret a positive coefficient of IQ in a schooling equation containing family-income or SES-level variables as implying that families accentuate human capital inequality. They assume that all the costs are borne and all the decisions about schooling are made by parents. But much of the cost of schooling is actually borne by the child and by the society at large. The question is what their schooling would have been without parental intervention. This is answered by the overall (between individuals) IQ coefficient. The fact that the within-families coefficient is lower I interpret as evidence of attenuation. Families do not go all the way with the IQ differences among their children as far as their investment in human capital decisions is concerned. This raises questions about the achievement models discussed in the earlier sections. They all assume that the "true" equation is the same within and between families. This assumption provides the connection between the within- and between-families parameters. Relaxing it will reduce the identification of several of the models examined and throw doubt on the attempts to estimate "complete" models, combining both the within- and between-covariance matrices.

[15] These computations are based on the Navy subsample of the NRC twins for which GCT scores were available. There are problems with interpreting such "late" scores as reflecting early IQ, and the reported results may be largely due to errors of measurement (though it would require an error rate of about 50 percent in the GCT to rationalize the DZ results). They are included here nevertheless, to indicate the general agreement of various bodies of data on this point.

least relative to a socialist state, which would take the children and allocate resources proportionally to natural ability so as to maximize the overall social product, families may actually reduce rather than accentuate potential inequality.

References

Becker, Gary S. *Human Capital.* New York: Columbia Univ. Press, 1964.
Becker, Gary S., and Tomes, Nigel. "Child Endowments and the Quantity and Quality of Children." *J.P.E.* 84, no. 4, pt. 2 (August 1976): S143–S162.
Behrman, J., and Taubman, P. "Intergenerational Transmission of Income and Wealth." *A.E.R.* 66 (May 1976): 436–40.
Behrman, J.; Taubman, P.; Wales, T.; and Hrubec, Z. "Inter- and Intragenerational Determination of Socioeconomic Success with Special Reference to Genetic Endowment and Family and Other Environment." Mimeographed. Univ. Pennsylvania, 1977.
Bielby, W. T.; Hauser, R. M.; and Featherman, D. C. "Response Errors of Non-Black Males in Models of the Stratification Process." In *Latent Variables in Socio-economic Models,* edited by D. J. Aigner and A. S. Goldberger. Amsterdam: North-Holland, 1977.
Bishop, John. "Biases in the Measurement of the Productivity Benefits of Human Capital Investments." Discussion Paper no. 223:74, Univ. Wisconsin, Inst. Res. Poverty, 1974.
Brittain, John A. *The Inheritance of Economic Status.* Washington: Brookings Inst., 1977.
Chamberlain, G. "Are Brothers as Good as Twins?" In *Kinometrics: The Determinants of Socio-economic Success within and between Families,* edited by Paul Taubman. Amsterdam: North-Holland, 1977 (*a*)
———. "An Instrumental Variable Interpretation of Identification in Variance-Components and MIMIC Models." In *Kinometrics: The Determinants of Socio-economic Success within and between Families,* edited by Paul Taubman. Amsterdam: North-Holland, 1977. (*b*)
———. "Identification in Variance Components Models." *Econometrica,* 1979, in press.
Chamberlain, G., and Griliches, Z. "Unobservables with a Variance-Components Structure: Ability, Schooling and the Economic Success of Brothers." *Internat. Econ. Rev.* 16 (June 1975): 422–49.
———. "More on Brothers." In *Kinometrics: The Determinants of Socio-economic Success within and between Families,* edited by Paul Taubman. Amsterdam: North-Holland, 1977.
Conlisk, John. "Can Equalization of Opportunity Reduce Social Mobility?" *A.E.R.* 64 (March 1974): 80–90.
Corcoran, Mary; Jencks, C.; and Olneck, M. "The Effects of Family Background on Earnings." *A.E.R. Papers and Proc.* 66 (May 1976): 430–35.
Goldberger, A. S. "The Genetic Determination of Income." Univ. Wisconsin SSRI Workshop Series 7707, 1977.
Gorseline, D. W. *The Effect of Schooling upon Income.* Bloomington: Indiana Univ. Press, 1932.
Griliches, Zvi. "Estimating the Returns to Schooling: Some Econometric Problems." *Econometrica* 45 (January 1977): 1–22.
Griliches, Z.; Hall, B.; and Hausman, J. "Missing Data and Self-Selection in Large Panels." *Annales de L'INSEE* 30–31 (1978): 137–76.

Griliches, Z., and Ringstad, V. *Economies of Scale and the Form of the Production Function.* Amsterdam: North-Holland, 1971.

Hauser, Robert M., and Daymont, Thomas N. "Schooling, Ability and Earnings: Cross-sectional Findings 8 to 14 Years after High School Graduation." *Soc. Educ.* 50 (July 1977): 182–206.

Heckman, James J. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables, and a Simple Estimator for Such Models." *Ann. Econ. and Soc. Measurement* 5 (Fall 1976): 475–92.

Jencks, C., and Brown, M. "Genes and Social Stratification: A Methodological Exploration with Illustrative Data." In *Kinometrics: The Determinants of Socio-economic Success within and between Families,* edited by Paul Taubman. Amsterdam: North-Holland, 1977.

Jencks, C.; Smith, M.; Acland, H.; Bane, M. J.; Cohen, D.; Gintis, H.; Heyns, B.; and Michelson, S. *Inequality: A Reassessment of the Effect of Family and Schooling in America.* New York: Basic, 1972.

Johnson, Harry G. "The Alternative before Us." *J.P.E.* 80, no. 3, pt. 2 (May/June 1972): S280–S289.

Lindert, P. A. "Family Inputs and Inequality among Children." Discussion Paper no. 218–74, Univ. Wisconsin, Inst. Res. Poverty, 1974.

Olneck, M. "The Effects of Education on Occupational Status and Earnings." Discussion Paper no. 358-76, Univ. Wisconsin, Inst. Res. Poverty, 1976.

———. "On the Use of Sibling Data to Estimate the Effects of Family Background, Cognitive Skills, and Schooling: Results from the Kalamazoo Brothers Study." In *Kinometrics: The Determinants of Socio-economic Success within and between Families,* edited by Paul Taubman. Amsterdam: North-Holland, 1977.

Taubman, Paul. "The Determinants of Earnings: Genetics, Family, and Other Environments: A Study of White Male Twins." *A.E.R.* 66 (December 1976): 858–70.

Zajonc, R. B. "Family Configuration and Intelligence." *Science* 192 (April 16, 1976): 227–36.