KINOMETRICS: DETERMINANTS OF SOCIOECONOMIC SUCCESS WITHIN AND BETWEEN FAMILIES

CONTRIBUTIONS
TO
ECONOMIC ANALYSIS

116

Honorary Editor

J. TINBERGEN

Editors

D.W. JORGENSON

J. WAELBROECK

NORTH-HOLLAND PUBLISHING COMPANY
AMSTERDAM · NEW YORK OXFORD

KINOMETRICS: DETERMINANTS OF SOCIOECONOMIC SUCCESS WITHIN AND BETWEEN FAMILIES

Editor

PAUL TAUBMAN

Professor of Economics
University of Pennsylvania

1977

NORTH-HOLLAND PUBLISHING COMPANY
AMSTERDAM · NEW YORK · OXFORD

Library of Congress Catalog Card Number: 77-15060

ISBN North-Holland for this series: 0 7204 3100 X
ISBN North-Holland for this volume: 0 444 85018 X

Publishers:

NORTH-HOLLAND PUBLISHING COMPANY
AMSTERDAM · NEW YORK · OXFORD

Sole distributors for the U.S.A. and Canada:

ELSEVIER NORTH-HOLLAND INC.,
52 VANDERBILT AVENUE, NEW YORK, N.Y. 10017

Library of Congress Cataloging in Publication Data
Main entry under title:

Kinometrics.

(Contributions to economic analysis ; 116)
1. Eugenics--Addresses, essays, lectures.
2. Success--Addresses, essays, lectures. 3. Family--Economic aspects--Addresses, essays, lectures. I. Taubman, Paul, 1939- II. Series.
HQ755.35.K56 301.42 77-15060
ISBN 0-444-85018-X

PRINTED IN THE NETHERLANDS

Introduction to the series

This series consists of a number of hitherto unpublished studies, which are introduced by the editors in the belief that they represent fresh contributions to economic science.

The term 'economic analysis' as used in the title of the series has been adopted because it covers both the activities of the theoretical economist and the research worker.

Although the analytical methods used by the various contributors are not the same, they are nevertheless conditioned by the common origin of their studies, namely theoretical problems encountered in practical research. Since for this reason, business cycle research and national accounting, research work on behalf of economic policy, and problems of planning are the main sources of the subjects dealt with, they necessarily determine the manner of approach adopted by the authors. Their methods tend to be 'practical' in the sense of not being too far remote from application to actual economic conditions. In addition they are quantitative rather than qualitative.

It is the hope of the editors that the publication of these studies will help to stimulate the exchange of scientific information and to reinforce international cooperation in the field of economics.

The Editors

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

P. Taubman

Kinometrics is the methodology that allows one to use samples with genetically linked relatives or kin to study the roles of genetics and family environment in a variety of areas including educational attainment, cognitive skills, occupational success, and earnings attainment. There are two different but related concerns in kinometrics. One of these is controlling for unmeasured or unobserved variables when estimating the relationship between measured variables. A good example of this is controlling for "ability" and "motivation" when estimating the effects of education on earnings. The second problem is measuring the combined and separate effects of the unobserved genetic and family environment variables.

The letter inviting people to participate in this conference indicated that the major purposes were to advance this methodology and to develop a body of substantive results on the determinants of educational and occupational success. It was also hoped that the exchange of ideas and results would stimulate the individual participants to expand their own work. I think the reader of this volume will find that the first two goals were fulfilled. Editors and readers of journals will soon find that the third goal has become a reality.

The book begins with a survey of the literature on family effects by A. Leibowitz and indicates a list of unresolved questions. These are addressed in the following five papers by Behrman, Taubman and Wales (BTW), Griliches and Chamberlain (GC), Jencks and Brown (JB), Olneck, and Sewell and Hauser (SH), each of which is based on a different and unfortunately nonrandom sample. The next section contains three papers. The first by Chamberlain examines the statistical methodology of combined latent variable, variance components models which are the basis of much of the work in the BTW and GC papers and are employed by SH. Briefly a latent variable model is one in which the same unobserved variable appears in or is related to several observed variables. A variance component model is one in which the variance of a variable is apportioned into several components all of which are unobserved. For example, in the

GC piece, the components are family and individual specific elements while in the BTW paper, the components are specific environment, family environment and genetic endowments. The basic questions that Chamberlain poses are how we determine which parameters we can estimate or identify in such systems, and how do we obtain additional information which will allow us to identify additional parameters. An oversimplified summary of this paper is that information on relatives allow us to control for unobserved variables in a latent variable framework provided the model can be specified in an appropriate fashion. In his second paper, Chamberlain also examines whether there are additional advantages in having information on twins rather than siblings. This issue will be examined below.

The paper by Goldberger contains a lengthy critique of the statistical methodology used in papers by BTW and JB. His major point is that some of the parameter estimates are obtained by imposing restrictions (making assumptions) which he considers arbitrary. In other words, Goldberger questions the assumptions made to identify their models. The reader unfamiliar with the problems of identification in the latent variable model may find Goldberger's discussion on pages , of particular versions of models in the Behrman, Taubman and Wales paper a useful companion to the more general discussion in the Chamberlain piece.

All the papers revolve around the question of how one can control for and measure the separate and/or combined contributions of unmeasured aspects of genetic endowments and family environment. Suppose that our model is

1) $$Y = \alpha X + G + N + u$$

where Y and X are observed variables such as earnings and schooling, G is an unobserved index of genetic endowments, N is an unobserved index of family and other systematic environments, and u is the unobserved and unsystematic part of the environment. With no loss in generality, we have standardized G and N so their coefficients are 1. We wish to obtain unbiased estimates of α and of $b^2\sigma^2_G$, $c^2\sigma^2_N$, σ_{GN}, σ_{GX}, σ_{XN} , and σ^2_u.

As is well known if b is nonzero and G is correlated with X, ordinary least squares of Y on X will yield biased estimates of α . Similar comments apply to X and N. Data on relatives can be used to control for G and N because kin share common or correlated genetics and family environments. The simpliest technique to exploit kin data is to use OLS with "within kin" observations. Suppose there are only two members of each family. Assign randomly the subscript of 1 or 2 to each member and then calculate the within-family difference as, e.g. $Y_1 - Y_2 = \Delta Y$. We then have

2) $$\Delta Y = \alpha\Delta X + b\Delta G + c\Delta N + \Delta u$$

If ΔG and ΔN are not correlated with ΔX, OLS of ΔY or ΔX will yield unbiased estimates of α — provided that ΔX is not equal to zero and is measured without error. N is defined to include common and specific elements which are the same for the twins or kin and can be represented by the average value in a family. The specific systematic environment is the difference between kin's actual environment and this average. Random events or nonsystematic environment is included in u. For identical twins ΔG is zero while ΔN will include only the twin specific part of systematic environment. For fraternal twins and siblings, neither ΔG nor ΔN is zero. As long as ΔN is uncorrelated with ΔX; within pair equations of ΔY on ΔX for identical twins will be unbiased, however, the equations for siblings and fraternal twins will still be biased as long as genetics matter. While the latter equations contain partial controls for G and N, these equations need not yield less biased estimates of α than is obtained from individual data.

Olneck presents within-pair equations for siblings, while BTW estimate such equations separately for identical and for fraternal twins. In both samples the subjects were drawn from about the same age cohort and were surveyed at about the same date. From a genetic viewpoint, siblings and fraternal twins are the same, though one might suspect that the correlation in environment for fraternal twins is greater. Thus it is of some interest that the Olneck and BTW results for sibs and fraternal twins respectively yield very similar regressions. Comparing the results for individuals and within families, both papers suggest a marked reduction in the coefficient on years of schooling in the equation for earnings. In addition, the BTW equations using within pair data for identical twins yield even lower coefficients on schooling which suggests that it is important to control for genetics.

There are a number of previous studies which have included IQ in an earnings equation. Thus it is interesting to note that when Olneck incorporates IQ into his within-sib equation relating earnings to schooling, he finds a further marked reduction in the coefficient of schooling. Thus differences in IQ remaining after eliminating common environment and differences in schooling are related to differences in earnings. Moreover, it is important to control for these IQ differences, which can arise from either genetic endowments or specific environment, when estimating the effects of schooling on earnings. However, Olneck's within-sib estimate of the coefficient on education after controlling for IQ is still above that of BTW in their within pair equation for earnings for identical twins. This suggests that abilities in addition to IQ need to be controlled. JB also found that differences in IQ are important in educational attainment equations calculated within pairs or using latent variable techniques.

GC also use data on siblings. While much of their analysis is based on a combined latent variable/variance components model which is described below, they report that within-sib equations for earnings expected at age 30 are the same as those calculated for individuals. It is not obvious why their sample gives such different results than do those of BTW and Olneck, nor is it obvious which set of results are more valid. However, it is possible that the differences arise because GC are studying men born more recently and with far fewer years of labor force experience than in the other two samples.

Even the equations for identical twins in the BTW paper, will be biased if there is measurement error in years of schooling or if ΔN is correlated with ΔX. In Appendix A of the BTW paper, there are estimates of what the coefficient of schooling would be if there were various percentages of measurement error in the schooling variable. The substantial bias found by BTW between the individual and within pair equations can only be eliminated if it is assumed that nearly all the observed within-pair variation in identical twins is due to measurement error.

Assessing the importance of ΔN leads us to the latent variable models which are employed by GC, BTW, SH and JB and analyzed in detail in both the Chamberlain and Goldberger papers. A latent variable is an unmeasured variable which appears in several equations. Under some circumstances, it is possible to control for and estimate the contribution of a latent variable to the variance of an observed variable. The first paper by Chamberlain investigates some of the conditions under which it is possible to identify parameters such as α in equation 1 and the contributions of the unmeasured variables. The paper by GC applies this technique to a model in which the observable variables include several measures of cognitive skills, years of schooling and expectations of earnings at age 30 which are held 5 to 10 years earlier. In their work and in Chamberlain's separate contribution, the focus is on a model in which the latent variable has a family (common) and individual component. That is, if there were only equations for Y and Z, if A is the unmeasured variable, if X is a vector of exogenous measured variable, and if i is the individual and j his family, the reduced form of the model would be:

3) $$Y_{ij} = \alpha X_{ij} + bA_{ij} + u_{ij}$$

4) $$Z_{ij} = cX_{ij} + dA_{ij} + v_{ij}$$

5) $$A_{ij} = F_j + w_{ij}$$

Here F_j is the family component.

While GC do not try to separate F into its genetic and environmental components, it is instructive to do so here. Let $\bar{G}_j$ be the average value of genetic endowments in a family and t_{ij} be the difference between G_{ij} and $\bar{G}_{ij}$. Similarly let $\bar{N}_j$ be the common component of N and r_{ij} be the difference between N_{ij} and $\bar{N}_{ij}$. Then we can write

5a) $$A_{ij} = \bar{G}_j + \bar{N}_j + t_{ij} + r_{ij}$$

where $F_j = \bar{G}_j + \bar{N}_j$ and $w_{ij} = r_{ij} + t_{ij}$.

The model used by BTW, and explored briefly in JB, differs in several ways from the GC piece. First, BTW implicitly assume that:

5b) $$A_{ij} = G_{ij} + \bar{N}_j + r_{ij}$$

That is they divide N into within and between family components but do not so divide G. Second, in their analysis, BTW at times impose the restriction that only $A_{ij} - r_{ij}$ enters into equations 3 and 4. If there were no genetic elements, this latter assumption in Griliches and Chamberlain's model would be equivalent to assuming that only F_j and not A_{ij} appears in equations 3 and 4. In the BTW framework this means that common and specific environments can have different effects on Y and Z. Moreover BTW do not impose the restriction that G_{ij} and N_j enter all the equations via A_{ij}, that is their coefficients need not be proportional in the various equations. By making these last two "asymmetric" restrictions, BTW are able to identify the estimates of α in their model. If these asymmetric restrictions are appropriate, it is possible to do more with samples of twins than with samples of brothers. However, if these restrictions are not valid, Chamberlain's theorem in his second paper, which indicates twins are no more valuable than sibs to estimate α, holds.

In the BTW paper, the estimates of the parameters on schooling and the other observed variables obtained from the latent variable technique are nearly identical to those obtained from the within-pair equations for identical twins. This comparability holds up under a variety of restrictions on the genetic and environmental variables including a version of the model in which there are no genetic components to A_{ij}. Thus BTW find that not controlling for genetic endowments and family environment or for "ability" leads to a substantial bias.

On the other hand, GC find in their latent variable that not controlling for family effects or for "ability" leads to only a small bias. Moreover, they find that the omitted variable that correlates highly with expected earnings has a low correlation with IQ.

As noted earlier, the differences in results for BTW and GC may reflect the years of work experience and age of the people in the samples used. It is also possible, however, that the differences reflect the structure of the models. In GC the available information from which the latent variables are constructed is heavily dependent on cognitive skills. The BTW model, which involves schooling and earnings at several points in the life cycle, would seem to encompass more skills. However, in their work labels such as cognitive ability cannot be used since for none of these skills are measures available. It is possible that future work will help resolve these issues. Both GC and JB find that IQ and schooling load heavily but not exclusively on the same variables.

Much previous work has tried to control for family factors by including observed variables such as parental education and number of siblings. The results available in all but the GC study suggest that the readily available measures do not adequately measure the contribution of, nor estimate the bias from not controlling for genetics and common environment. SH, for example, find that parental education, father's occupation, broken family, and number of sibs account for about half the cross-sib correlation in educational attainment.

Thus far we have been discussing the control of family factors. Now let us examine the measurement of their effects. (The label family is somewhat of a misnomer since sibs may share common environments in the military and other nonfamily settings, and because even in the family, sibs may be treated differently or be exposed to different environments.)

The family effect is based on the cross-sib covariances or correlations. On all variables examined in this book, the correlation is always higher for identical than for fraternal twins. The fraternal twin and ordinary brothers cross-sib correlations are quite close though there is some question as to the comparability when only sibs close in age are used. (See JB who discuss the results from other samples and compare Olneck and BTW.) The few samples on identical twins yield fairly comparable estimates on cross-sib correlations and for education. The NAS-NRC twin samples in the BTW paper suggests that the family effects may account for about 75% of the variance in education and about 50% of the variance in earnings (late in the life cycle).

As explained in the BTW and JB papers, strong assumptions have to be made in the so-called classical twin analysis of variance model to identify the separate contributions of genetics and common environment to these family effects. JB, who test several of these assumptions such as there is no more interaction among MZ than DZ twins generally are not able to reject the usual assumptions made. BTW show that more of these assumptions can be tested in the combined latent variable/variance component models. However, one crucial assumption that neither they nor JB can test is that the cross-sib correlation in unmeasured environments is the same for identical and fraternal twins. As Goldberger cogently argues, it is possible to explain all of the increase in the cross-sib correlation of identical twins over fraternal twins on the assumption of greater similarity of treatment or environment of the identical twins. That is, it is always possible to argue that there are no genetic effects. There is some weak evidence in the papers that bear on this particular assertion. First, when BTW fit a pure environment model to the data, using the latent variable technique, they find it fits less well than a mixed genetics/environmental model. Second, if identical twins are treated more alike than fraternal twins, it would also seem that fraternal twins would be treated more alike than ordinary sibs who would be raised in somewhat different family and social environments. Yet a comparison of the brothers in Olneck and the fraternal twins in BTW indicate little difference in results in cross-sib correlations for education and earnings. However, Olneck, who has investigated this issue in more detail, reaches ambiguous conclusions on variation in cross-sib correlations by age.

It is important to remember that the debate over whether identical twins have a more highly correlated environment is a debate solely over the division of the so-called family effect. Several of the papers and much of the discussion of the conference revolved around whether or not this additional information was useful. Being too close to this debate to summarize it fairly, the reader is referred to the papers by Goldberger, BTW and JB.

But it is important to remind the reader that with the exception of the piece by GC, the empirical papers indicate that it is very important to control for family effects when estimating schooling, occupational status, and earnings equations. It appears that within pair equations estimated by OLS for identical twins yield results on the observed variables quite comparable to those obtained from latent variable techniques. Further, the latent variable techniques appear to be robust as to these same coefficients regardless of the assumptions made about genetics and environment.

The reader of these papers will be rewarded with a wide variety of results and with several methodologies which can be quite powerful in the appropriate circumstances. It seems likely that the method can be extended to incorporate more dependent variables or indicators. Some prime candidates may be health and fertility. It is possible to include additional categories of kin to refine further the estimates of genetic and family environmental effects. Also the technique will indicate the extent to which any set of measured variables, which can be collected in random and nonkin samples, represents the full family effects.

The people who were the formal discussants at the conference were invited to submit their comments for publication. Several discussants, however, who felt the authors in substantial revisions had responded fully to these comments chose not to publish their comments.

Finally, let me take the opportunity to thank all of those involved in the conference for their participation and cooperation. I particularly wish to thank Jere Behrman, Arthur Goldberger, and Zvi Griliches for their suggestions on organizing the conference. The conference was funded by the National Science Foundation by a grant to the Mathematical Social Science Board. I would like to thank both this Board and especially the economics representative, Marc Nerlove, for their generous support.

This material was prepared with the support of the National Science Foundation Grant # SOC-70-02316 A04, formerly GS-3256. Any opinions, findings or conclusions or results are those of the authors and do not necessarily reflect the view of the National Science Foundation.

And special thanks to ETC Business Services for their assistance in preparing the camera-ready copy of this volume.

CHAPTER 2

FAMILY BACKGROUND AND ECONOMIC SUCCESS: A REVIEW OF THE EVIDENCE*

Arleen Leibowitz
Law and Economics Center and
Department of Family Medicine
University of Miami

In a 1970 survey of the distribution of labor incomes, Jacob Mincer contrasted the human capital approach to the analysis of personal income distribution to the "traditional approaches centering on differences in opportunity, ability and chance." It then appeared that the analysis of income generation based on human capital theory had supplanted the rather ad hoc theories based on "opportunity, ability and chance."[1]

The human capital analysis was appealing not only because it put the decision to invest in schooling on the same rational basis as the decision to invest in other capital, but also because it was formulated in a way that was empirically testable.

It was not long before there was a renewal of interest in opportunity, ability and chance and these variables were reintegrated into the theory and estimation of income determination. On an analytical level these factors were treated as shifting the return to schooling and schooling cost schedules and therefore schooling attainment (Becker, 1967). Empirically much effort in recent years has been directed to measuring the impact of ability and opportunity or family background on both schooling and income. First ability, and later background measures have been added to the human capital variables in earnings functions. The human capital variables have proved to be quite robust, retaining significant effects on earnings even when measures of ability and background are introduced into earnings equations. While the direct effect of ability and background variables on earnings has proved to be small relative to the effect of education, the effect of these variables on education itself seems to be substantial.

From the initial calculations of rates of return to schooling, it was recognized that the positive correlation of ability with years of schooling would lend an upward bias to the rate of return calculations (Becker, 1964). In the years following the publication of Human Capital several papers analyzed the relation between ability and schooling and their impacts on income. While Hause (1972) found an interaction between ability and schooling at high levels of both variables, Griliches and Mason

*The support of the National Center for Health Services Research, through Grant No. HS01623 is gratefully acknowledged.

(1972) found that over-all the lack of the mental ability variables exerted very little bias to the schooling coefficient. In the presence of an index of mental ability, the schooling coefficient was 7% to 15% smaller than in an equation containing no measure of ability.

The "traditional" approaches stressed that not only characteristics of the individual, such as ability and schooling, but also characteristics of the individual's family could influence earnings. The empirical estimation was next extended to include the influence of background variables on education, occupation and earnings. In sorting out the impact of ability on earnings, there was general agreement on what the measure of ability should be - since only one kind of ability measure was generally available - and that was usually the score of an I.Q. test or some other test which correlates well with I.Q. - such as AFQT.[2] In contrast, there is little concensus on the appropriate background measures. Parental education, occupation, earnings, residence, time inputs to children and college quality are among the measures which have been used. Studies using data on brothers point to the possibility that there are additional family background variables which affect earnings and occupation but they're not the variables we've looked at so far.

Part I reviews some studies that trace economic achievement to differential human capital and background variables. Part II reviews studies which utilize data on brothers to analyze the elements in family background for which we have no ready direct measures. In Part III the points of agreement and disagreement among these studies will be summarized.

I. Economic Success and Measured Family Background

The influence of family variables on economic achievement can either be direct - for example, earning higher income because one's father had high income - or indirect - earning high income because one's family purchased schooling. The analysis of these background factors and the manner in which they influence income, occupation and intermediate levels of achievement such as schooling has absorbed a great deal of attention in recent years.

Most of the studies have in common a model, composed of one or more of the following set of equations, which are assumed to be block recursive and are estimated by OLS.[3]

1) $A = a_1X + eG + u_1$

2) $S = a_2X + b_2A + u_2$

3) $O = a_3X + b_3A + c_3S + u_3$

4) $Y = a_4X + b_4A + c_4S + dO + u_4$

A set of socio-economic background variables, X, and genetic inheritance, G, are exogenous, and an early measure of ability, A, depends on both. The quantity of schooling attained, S, is determined by various background measures, X, ability and a random term u_2. Occupational status, O, is a function of background, ability, years of schooling, and a random term, u_3. Income, Y, in turn, depends on X, A, S, occupation and a random term.

To determine the total effect of background on schooling, for example, we must consider not only the coefficient, a_2, but also the indirect effect of background on ability and, thus, on schooling.

One of the early examples of this type of model is presented by Duncan (1968). He used data from the 1962 Occupational Changes in a Generation survey and a 1964 CPS update to estimate a path model for the achievement of white men, aged 25-34 in 1964. While 20 of the necessary correlations could be computed from these sources, the correlation between early I.Q. and educational attainment, between early and late I.Q. and between I.Q. and family size had to be estimated from other sources. Five other necessary correlations were inferred from the model.

The model estimated had four dependent variables: adult I.Q., educational attainment, occupational achievement and earnings. Educational attainment was assumed to depend on early I.Q., number of siblings, father's occupation and education. Adult I.Q. depended only on childhood I.Q. and years of schooling, while occupational achievement was dependent on adult I.Q., father's occupation, number of siblings and schooling. Earnings were assumed to directly depend on all the other variables in the model except early I.Q.

The estimation indicated that the rank order of relative importance of the four family background variables was the same in explaining schooling, 1964 occupation and earnings.[4] Early I.Q. accounted for 38% of the explained variance in schooling, 30% of the explained variance in occupation and 45% of the explained variance in earnings. It must be noted, however, that the proportion of variance explained and the size of the standardized regression coefficients is larger, the closer is the outcome measure to the beginning of the causal chain. The relative importance of ability vis-a-vis background measures may be due to the lack of data on other relevant background variables, such as family income or mother's education.

The most important determinant of occupational achievement is years of schooling, which has a direct path coefficient of .50. Duncan notes that the direct influence of number of siblings, father's education and occupation is equalled by their effect transmitted through educational attainment.

Duncan finds that occupation has the greatest path coefficient among the determinants of earnings (p=.26). Adult intelligence has a path coefficient of .13 and education of .09. But the indirect effect of education via occupation is .13, which exceeds the direct effect. Number of siblings and father's occupation were not significantly directly related to earnings, and the path coefficient of father's education was .03. However, these variables do affect earnings indirectly, through their relationship to schooling and occupation. This implies that earnings are primarily a function of individual characteristics, not a pure rent on family background, but that family background is related to the characteristics that do determine earnings.

It has been argued in Bowles (1972) and Bowles and Nelson (1974) that the lack of significance of family background variables, found by Duncan, results because there is little measurement error in I.Q. and years of schooling, while the parental status variables are derived from retrospective data, and, therefore, are subject to greater error of measurement. This would bias their coefficients toward zero in estimating earnings functions, while the impact of I.Q. and years of schooling is likely to be overestimated because they are correlated with the unmeasured background elements.

Bowles and Nelson, using observations on four age cohorts of non-farm, non-negro males from the 1962 Occupational Changes in a Generation Survey, constructed a variable representing parental income using independent information on its relationship to the reported background variables. They also corrected the observed variables for measurement error.

In estimating the schooling equation, they find both childhood I.Q. and a group of socioeconomic background variables are quite significantly related to educational attainment for all four age cohorts - men aged 25-34, 35-44, 45-54, and 55-64. I.Q., socioeconomic background[5] and years of schooling are generally significant predictors of occupational status and income in the same four age groups. Bowles and Nelson note that the normalized regression coefficients of their "corrected" socioeconomic background variables on schooling, occupation and income, exceed the coefficients of I.Q., in contrast with the findings of Duncan. The effect of the corrections is to raise the standardized regression coefficients. Bowles and Nelson conclude on the basis of these "corrected" data, that I.Q. itself is not the major vehicle whereby parents are able to transmit economic success to their children. Rather, they hypothesize that the genesis of the correlation between children's and parents' achievement is "in aspects of family life related to socioeconomic status and in the effects of socio-economic background operating both directly on economic success, and indirectly via the medium of inequalities in educational attainments."

In spite of the crudeness of the corrections, the results obtained by Bowles and Nelson do correspond to studies where the background data are contemporaneous and fairly accurately measured. In their study of 1957 Wisconsin high school graduates, Sewell and Hauser (1975) have background information not derived from retrospective data. Measures of parental income were obtained from Wisconsin income tax returns for the year in which the subject was a high school senior (1957) and the three following years. This probably is a good index of the ability to afford college and provide advantages in the working world. Own income was quite accurately measured since it was derived from Social Security records. Sewell and Hauser find that ability has a large direct effect on schooling, but that in determining earnings, both schooling and ability are dominated by own occupational status (related to schooling) and parental income.

Equations representing all four types of outcomes have been estimated by Sewell and Hauser (1975). The initial data were derived from responses by 1957 Wisconsin high school seniors to a questionnaire regarding their socioeconomic background and educational plans. This information was supplemented with data from a 1964 follow-up which determined the students' post-high school educational and occupational attainments. Information on post-high school accomplishments was obtained from the parents of the seniors for whom followup data was solicited. For these men, earnings in 1965, 1966 and 1967 were estimated from Social Security records after correcting for the effects of multiple job holding and the ceiling on earnings subject to Social Security taxes. The 4388 men on whom data were obtained from the 1964 follow-up survey were found to represent fairly closely the original respondents to the 1957 survey.

In estimating a version of equation (1), on 2069 observations where the respondent was employed, not enrolled in school and non-farm, Sewell and Hauser find that son's measured ability is related to several measures of socioeconomic background: father's and mother's educational attainment, father's occupation and parents' average income. These variables also have about equal path coefficients in the direct determination of educational attainment while ability itself has a large direct effect on years of schooling.

Occupational status seven years after high school graduation as measured by Duncan's SEI is most strongly related to educational attainment with each year of post high school education worth about six points on the Duncan index. The only background variable with a direct effect on occupational status is father's occupational status.

Among the background variables affecting income, only parents' income had a direct influence on son's earnings. Neither father's education nor occupation had a significant independent effect on son's earnings if father's income is included in the equation. However, these variables do affect income indirectly through the ability and schooling variables, but these effects may be modest. They conclude that only 20% of the zero order correlation between ability and income need be attributed to socioeconomic background, as they measured it, while measured SES variables alone account for 15-18% of the variance in schooling.

Sewell and Hauser find that in determining earnings "ability and educational attainment have modest and roughly equal effects, while the effects of occupational status and parental income. . . .are about twice as large as those of ability and educational attainment."[5] These results contrast with other studies in finding education and ability equally important, and that they are surpassed by parental income. Education may appear to be a less important variable in determining earnings here than in other studies because an index of occupation is also included in the equation. When the equation is estimated without the occupation measure, schooling and parental income are about equally imporant, while ability has a Beta coefficient only 60% as large. There may be an upward bias to the ability coefficient due to ability's relationship with market experience, since Sewell and Hauser find that both the date of college entrance and the final attainment level are positively correlated with ability. Thus, holding years of schooling and age constant, greater ability is positively related to post-school labor market experience. It will be interesting to see as the data collection continues if schooling and ability are equally important in determining earnings at age 50.

Part of the explanation for the importance of parental income in the earnings function may lie in the greater market experience of children from richer families. Labor market experience may vary with family income because children from poor families may take more years to complete their degrees because of the need to work to finance their schooling.[7] Sewell and Hauser do examine several correlates of family income and reject the hypothesis that the impact of father's income on son's earnings is due to the place of residence, place of origin, military experience, marital status or labor force experience. College quality may also be a link between father's and son's earnings. They conclude that "slightly more than half of the variance in earnings among colleges in (their) sample was explained by the fact that colleges select or recruit men with varying prospects for earnings.[8]

In this sample, data on the mother's characteristics were available in addition to the data on father's which most studies rely on as indices of family environment. Variables measuring the mother's education and income significantly improve the amount of variance explained in the ability, education and occupation equations. Including mother's income with father's income increased the amount of explained variance in son's earnings. This provides some support for the "production" explanation for the importance of background variables as opposed to the transmission of social class approach. An independent effect of mother's income on son's earnings is consistent with the hypothesis that greater parental income and productivity lead to more investments which increase the son's earnings capacity.

Can the effect of parents' income on earnings be traced to other social variables? Sewell and Hauser find no evidence to support this contention. While measures of teachers' and parents' encouragement of college plans, high school grades, own and friends' plans for college and occupational aspiration were good predictors of educational attainment and occupational level, none was significantly related to income.

Sewell and Hauser conclude that much of the influence of social background on earnings, occupation and schooling is "due to the superior cognitive and motivational environment" provided in higher SES homes. However, they see evidence of "ascriptive elements" permitting direct transfer of occupational and economic status, as in the effect of parents' income on son's earnings.

The evidence reviewed so far is consistent with the statement that employers pay primarily for characteristics of the individual - his schooling or ability - and that most background variables affect earnings indirectly through their effects on earnings and ability. However, parental income does seem to have a direct impact on son's earnings. Perhaps parental income is related to other unmeasured dimensions of an individual's earning capacity for example, quality of schooling.

Another example of a recursive model estimated on individual data with contemporaneous data on family background is presented by Conlisk (1971). He estimates versions of equations 2 and 4 of the prototype model presented above using a sample of men from the Berkeley Guidance Study. In this data I.Q. and other variables were measured at various ages from infancy to age 18 while education and occupation data were collected at age 30, for about 75 children born in Berkeley, California in 1928-29. Average earnings from the 1960 Census were assigned to occupations in order to estimate earnings.

Education proves to be a significant predictor of earnings, with each year of schooling accounting for $500-$600 of the earnings proxy. However, I.Q. measured at various ages between 1 and 18 years in separate regressions, never has a coefficient which is significantly different from zero. Parents' income (observed at the time of the child's birth) is significantly related to the earnings proxy, but parents' average years of schooling was not. In spite of its lack of direct impact on estimated earnings, I.Q. does have an indirect impact through schooling. In regressions where years of schooling completed is the dependent variable, both I.Q. and parents' income are significant predictors, while average level of parental education is not. I.Q. measured at age 12-14 has a coefficient which is twice as great in explaining schooling as I.Q. measured at ages 1-5.

Wachtel (1975) analyzed the effect of both high school and college quality on later achievements using 1812 observations from the NBER - Thorndike sample. This sample consists of men who in 1943 were tested for the Army Air Force pilot and navigator programs. In addition to the 1943 test scores, family and personal background measures were collected at this time. Income, occupation and earnings, and other background data were collected from this sample in 1955, 1968-69, and 1970-71. High school quality measures were generated by Wachtel who matched high school names and locations to school districts in order to calculate measures of high school quality from the U.S. Office of Education's Biennial Survey of Education for 1936-1938. Matches were possible for 46% of the sample members. College quality was measured alternatively by per student expenditures and by the Gourman rating.

Wachtel found that measures of school quality at both the high school and college level are significant determinants of earnings. The path coefficient of college quality (measured by total direct expenditures for undergraduate and graduate schooling) is as large as those for ability and experience in the earnings function. It is exceeded by the path coefficients for years of schooling and log hours worked. Quality of high school has a direct path coefficient in the earnings function of about the same magnitude, but also has indirect effects due to its positive relationship with measured ability and college quality. There was little evidence of a direct effect of high school quality on years of completed schooling.

In Wachtel's ability equation, the coefficients on the other independent variables - number of siblings, mother's education and father's education - remain virtually unchanged when school quality is added. The same is true in the equations explaining years of schooling and income.

Wachtel attempts to test the hypothesis that school quality is merely a proxy for family wealth or community income by including in the regressions, 1939 median income in the state where the respondent attended high school. School quality retains its significant positive effect in the earnings functions, while neither school quality nor state income has a significant coefficient in the equations explaining ability or years of schooling. It may well be that per capita expenditures in a school district is a better proxy for family income than is median state income. It would be interesting to have a better measure of family income in this data set to determine if family income has an effect on earnings net of school quality.

Solmon (1973), analyzing the same data set, found that earnings were affected by two distinct aspects of college quality - faculty quality (as measured by average salaries) and student quality (as measured by average student test scores). The impact of quality of schooling appeared to rise with age, both for men with 12-16 years of schooling and for those with graduate training. For those with graduate training, quality of undergraduate college was not significantly related to earnings on first job, in 1955 or in 1969. While graduate school quality was not related to earnings in the first job, it was significantly related to earnings in 1955 and 1969. Introducing variables for father's and maternal grandfather's schooling and a rough index of father's occupation (indicating high, medium and low) reduced the size and significance of the quality variables slightly, but the quality variables remained significant predictors of 1969 earnings,.

Morgenstern (1973) investigated the effect of school quality on the earnings of blacks and whites with data from the 1968 urban problems survey, conducted by the Survey Research Center at the University of Michigan. The sample consisted of 842 black and 782 white male and female employees living in 15 largely northern cities and two suburbs.

Measures of school quality were constructed by assigning to each respondent the average value of per pupil expenditure, average teacher salaries and student-teacher ratio which applied to the state and decade in which the respondent attended primary school. In wage equations estimated separately for blacks and whites, wages are a positive function of schooling and experience, but are lower if the father's education level was low and if the respondent was a female. In spite of likely downward bias to the coefficients on school quality variables due to the use of state-wide measures, whereas the intra-state variation in quality is likely to be correlated with measured background variables, school quality proves to have a small direct effect on the wages of blacks.

Per pupil expenditures are significantly related to earnings for blacks, although the coefficient becomes not significantly different from zero when a North-South dummy variable is included. In none of the formulations do wages of whites vary with school quality variables.

In explaining educational attainment, per pupil expenditures are significant predictors for both races even holding constant region of birth. Using parental education as a proxy for socioeconomic background, Morgenstern found that mother's education is more strongly related to educational attainment of blacks, while father's schooling is a better predictor for whites. Thus, schooling attainment seems to be significantly related to socioeconomic background and school quality for both blacks and whites, while the direct effect of school quality on wages was seen only for blacks.

A production function approach to the significance of family background variables has been taken by Leibowitz (1974) and Lindert (1976). Leibowitz argued that the significance of maternal education in explaining childhood ability may be due to the fact that time inputs per child increase with maternal education. Direct measures of time inputs to children proved to be good predictors of childhood I.Q. scores in a sample of 821 high ability males known as the Terman Sample. Further support for a production interpretation comes from the greater effect of mother's education on I.Q. than of father's,[9] as one would expect since mothers spend more time with children than fathers. In explaining years of schooling, mothers' and fathers' education were about equally important, but time inputs to children had no direct effect apart from their influence on I.Q., which was significantly related to schooling attainment. Family income in childhood was not directly related to years of schooling. Greater numbers of siblings were associated with lower levels of completed schooling by age 30, but an estimation of schooling completed by age 40 showed no significant relationship with family size. Thus, larger numbers of siblings may cause schooling to take longer to complete (due to the strain on family resources or the obligation to put brothers and sisters through school), but the final level of schooling is independent of family size for this group of high ability students. The earnings function for this group showed a strong relationship between earnings and schooling and experience. However, childhood I.Q. and parental education were not directly related to earnings, once own level of schooling and experience were accounted for.

Parental income had a positive effect on son's earnings in 1940 (age 29) which diminished in size and significance by 1950, and had no significant effect on 1960 earnings. This decreasing impact of family income with age should be borne in mind when evaluating Sewell and Hauser's earnings functions which relate primarily to men under 30.

Although it was not possible to compute earnings functions for the 671 females in the Terman sample, the schooling equations are sufficiently different from those for males to warrant discussion. In contrast to the Terman males, schooling levels for females were not significantly related to ability, indicating greater equality of opportunity (independent of SES) for high ability boys than girls. The impact on schooling attainment of mothers' education exceeded that of fathers', while for boys, the effects of mothers' and fathers' attainment were equal. Girls with more siblings ended up with less schooling and did not "catch up" as boys did.

The effect of family size on schooling attainment and on measured I.Q. is open to two interpretations: that family size is correlated with unobserved parental variables (e.g. demand for high achievers) or that it reflects differential inputs to children. Peter Lindert, in a recent paper, "Sibling Position and Achievement" uses a cross section of intra-family sibling differences to test these two interpretations of the importance of family size and birth order. He finds the data consistant with the hypothesis that sibling position matters because of its postulated effect on parental time inputs to children.

It is well known that I.Q. declines with birth order and family size (Belmont and Marolla). If this results from lesser time inputs received by children who must share their parents with additional siblings, last born children should have an advantage over middle born children because they have exclusive claim on their parents' time for part of their lives, while middle children never do. This advantage may show up in measured ability, acquired schooling or earnings achieved.

Lindert uses data on 1087 siblings collected from interviews with 312 senior male employees of a New Jersey utility company in 1963 to trace birth order and family size effects on achievement. In explaining the variance in schooling attainment across all 1008 persons in the sample, mothers' and fathers' education both have coefficients which are highly significant and about equal in magnitude. Father's occupational status was positively related to schooling attainment, and a broken home was negatively related. First born children and those from small families obtained significantly more schooling than others, although the hypothesized disadvantage of middle children vis à vis last-borns did not pass significance tests. Age and sex were not significantly related to educational attainment.[10]

An index of males' occupational status was significantly related to fathers' education and occupation, but not consistently to mother's education. Men from broken homes had significantly lower occupational status and older men had a higher index rating. Again, men from small families and first born men had higher status, holding all other background variables constant. To overcome the problem that the interview respondents had higher achievement than their siblings, on average, and that the probability that any observation was an interview respondent was related to his family size (one half of children in two child families were interviewed, but only one third in three child families, etc.), regressions were run on intra-family differentials in all variables. Less of the variance is explained in these regressions (about 8%) than in the across family regressions (about 31%). It had been hypothesized that last borns get more time than middle children because they are the "only child" in their home for some time. The coefficients support this hypothesis, but are generally not significant. One comparison that was statistically significant was that last-borns among 4 or 5 child families do better relative to the first born than the 4th or 5th child in a family of six or more. This supports the contention that last borns have an advantage over middle children, as well as providing evidence for the family size effect.

When the set of sibling position variables was replaced by an index derived from time budget data on relative amounts of time spent with children in various size families and various birth orders, this index was significantly related to educational attainment across and within families and to occupational status across families. Lindert concludes that these results are consistent with the view that "sibling position matters because of its straightforward effects on family time and commodity inputs into children. . ." Yet the evidence is not conclusive since the variation in the time input index results largely from the greater time inputs received by first born children.

Lindert's analysis does point to a rationale for the importance of maternal education on economic outcomes in studies which do not hold constant sibling position. In these studies, maternal education may act as a proxy for differential inputs to children, since more educated mothers have fewer children and spend more time with each child.

The relative importance of education, intelligence and social background in explaining variations in personal income in Sweden has been analyzed by de Wolff and Van Slijpe. The data were provided by Husen and relate to a group of Swedish men who were third graders in the Malmo primary school in 1938 when the data collection began. At that time social class of the parents, and I.Q. were recorded. In 1963 the original group were resurveyed and queried about their 1963 taxed income, years of schooling, and occupation. Complete set of data were provided by 65%, or 545 men.

Both linear and semi-log earnings functions were estimated. In both formulations education is the most important explanatory variable, followed by social class, then childhood I.Q.

In de Wolff and Van Slijpe's experiments with the interactions among their independent variables, they found that intelligence and social class reinforce each other in generating earnings, and their combined effect, particularly at high levels of education, is much greater than implied by an additivity of effect. Their data indicate that the small number of highly intelligent children from low SES backgrounds who received high levels of schooling earned less than their counterparts from more advantaged backgrounds. However, even for a low level of social background and moderate level of intelligence, schooling has a positive effect on earnings.

II. Evidence of Unmeasured Elements of Family Background from Studies of Brothers.

The availability of data on brothers has allowed a kind of "natural experiment" which can provide some answers to the question whether the omission of relevant background variables correlated with educational level has biased estimates of the relationship between schooling and income. Using the data on brothers to run intra-pair regressions allows one to hold constant some elements of family environment, and data on monozygotic twins even allows one to hold constant genetic makeup.

To estimate the importance of genetics, home environments, and individual investments explaining variance on earnings, Paul Taubman used the NAS-NRC sample which contained data on white male twins born between 1917 and 1927, both of whom had served in the military. Taubman (1975) uses the fact that there are two types of twins - monozygotic - those who are derived from a single fertilized egg and are, therefore, genetically identical - and dizygotic - those who arose from two separate fertilized eggs. This allows one to assign part of the variance in earnings of dizygotic twins as due to genetic diversity by contrasting with monozygotic twins, for whom there is no genetic diversity within pairs. Some additional assumptions are also necessary. These include:

1) no assortive mating

2) no sex linked, dominant or recessive genes
3) a moderate degree of non negative correlation between genetics and environment.
4) equal environmental correlation for DZ and MZ brothers.

Calculating plausible ranges for the variables and assuming a correlation of .02 to .25 between genetics and environment, Taubman finds that genetics may account for between 11 and 33% of the variance in earnings at age 45-55 while common environment accounts for 11-17% and non-common environment 45-57%.

Goldberger points out that these bounds are not confidence intervals, and depend on the assumed values for the unidentified parameters in the model (1976, This volume, p. 8).

In accounting for the variance in schooling outcomes, genetic factors are related to 20-35%, common environment 29-32% and non-common environment 21-22%. It seems reasonable that genetics and common environment should account for more of the variation in schooling than earnings. But even in explaining earnings, "genetics" looms larger in Taubman's study than in analyses where specific "genetic" attributes, such as I.Q. (which is only partly genetically determined) are used. One might have expected that I.Q., which is not wholly genetically determined, would account for more variance than purely "genetic" traits. Taubman concludes from this disparity that the relevant genetic endowments encompass more than I.Q. or cognitive skills. This also seems reasonable, since other partly genetic variables, such as physical strength or height can influence earnings. While the Taubman paper establishes some ranges for the effects of genetics and environment under certain assumptions, it does not pinpoint the specific elements of the environment or of the genetic potential which affect income and schooling.

In reviewing correlations among brothers' earnings, Corcoran, Jencks and Olneck conclude that the intraclass correlation between brothers' earnings exceeds by a factor of two the percentage of variance in earnings explained by the measures of common background of the brothers. This implies that elements in the brothers' common background on which no measures were available were equally important in explaining earnings as those for which measures were available. Further, these unmeasured common variables are uncorrelated with schooling and ability. Corcoran et. al. hypothesize that these may be non-cognitive interpersonal skills valued by employers or a taste for non-money income versus money income.

Chamberlain and Griliches (1975) using data on the 1927 income of 156 pairs of brothers from Indiana (the Gorseline data) compare earnings functions for the total sample with estimates where each brother's characteristic is measured around his own family's mean. This eliminates from the relationship that part of family background and genetic inheritance which is common to the two brothers. Using OLS there was virtually no difference in the estimated marginal effect of schooling on income (a coefficient of .082 versus .08) between the two methods.

Estimating a structural model whose reduced form disturbances are connected by a common unobservable variable having a within and between group structure, the schooling coefficient in the maximum likelihood estimates of the income equation rises but the estimated coefficient from this recursive model (.084) is very close to that estimated by OLS. Chamberlain and Griliches comment "Although our estimator was carefully designed to detect omitted variables connecting and biasing the income and schooling relationships, we haven't found any."[11] What they do find evidence of, however, is an unobservable variable, which they name "ability," but which has little effect on years of schooling, while having considerable importance in the income and occupation equations. It is hard, therefore, to interpret this "ability" measure as I.Q. or family income, because other studies reviewed below have indicated the greater impact of I.Q. and parental income on a child's schooling than on his income. However, this "ability" variable behaves very much like Wachtel's measure of high school quality in affecting income, but not years of schooling. This common family factor might also be consistent with other family effects which are not represented by the variables we usually measure.

III. **Rates of Return or Rents: A Summary**

Table 1 outlines the results of the several studies reviewed above. The empirical analysis used many different sources of data, and several different estimation techniques. Most of the studies dealt with United States data, but their results were consistent with the study of Swedish men. A majority of the analyses used earnings data for the 1960's, but data for 1928, 1940, and 1950 told essentially the same story. The ages of the subjects were often in the 20's, but the same general pattern emerged for older men as well.

In spite of the variety of data and techiques, a pattern emerges from these studies. The pattern can be summarized as follows: Schooling attainment and other characteristics of the individual such as ability and labor market experience are the primary direct determinants of income and occupational status. While some background measures may be directly related to income and occupation, their impact is primarily via their influence on schooling attainment and measured ability. Family variables have a major impact on measured ability, and these variables together with ability explain much of the variance in schooling attainment. It thus appears that employers primarily pay for individual characteristics, such as schooling, ability or labor market experience. For the most part, earnings do not directly result from pure rents on family characteristics. However, schooling attainment and ability are strongly related to these background variables. If these studies do appear to come to a consensus about what is known, several studies using data on brothers come to similar conclusions about what is not known. They indicated that more of the variance in earnings is related to factors which are common within families than has been found using specific measures of parental education, occupation and income. Thus, elements of the family environment not crystallized in ability or schooling attainment may directly affect economic success - but the indices of family background we have been using do not measure all the relevant factors.

TABLE 1: OUTLINE OF MODELS OF ECONOMIC SUCCESS

Author	Sample	Dependent Variable	Related to	Not Related to	Variance Explained
Bowles and Nelson	1962 Occupational changes in Generation	1962 Income ages 25-34 35-44 45-54 55-64	Schooling 1962 Father's Occupation Education, Income Childhood I.Q.		15% 26%
		1962 Occupation ages 25-34 35-44 45-54 55-64	Schooling 1962 Father's Occupation Education, Income Childhood I.Q. (except age 25-34)		45% 58%
		1962 Schooling ages 25-34 35-44 45-54 55-64	Father's Occupation Education, Income Childhood I.Q.		41% 47%
Chamberlain and Griliches	Gorseline 156 pairs of brothers	1928 Income in Indiana	Schooling "Ability" Age		
		1928 Occupation	Schooling "Ability" Age		
		1928 Schooling	Age	"Ability"	

Author	Sample	Dependent Variable	Related to	Not Related to	Variance Explained
Conlisk	Berkeley Guidance Study 75 men	1960 Earnings (age c.30) (est from occupation data from 1960 Census)	Schooling (age 30) Parents' Income (1928-29)	I.Q. (ages 1-18) Parents Schooling Number of own children	43%
		Schooling (age 30)	I.Q. (ages 1-18) Parents' Income	Parents' Schooling	44%
de Wolff, Van Slijpe	1938 Students in Malmo, Sweden 545 men	1963 Taxed Income	Schooling Social Class Childhood I.Q. (I.Q. and Social Class Interaction)		33%
Duncan	1962 Occupational Changes in a Generation; 1964 CPS	1964 Earnings men aged 25-34	1964 Occupation Adult I.Q. Father's Education Education	Number of Siblings Father's Occupation	11%
		1964 Occupation	Education Father's Occupation Adult I.Q. Father's Education Number of Siblings		28%
		Education	Childhood I.Q. Father's Occupation Father's Education Number of Siblings		42%

Author	Sample	Dependent Variable	Related to	Not Related to	Variance Explained
Leibowitz	Terman Sample 700-784 obs. males 615 females	1940, 1950, 1960 Earnings (age 29, 39, 49)	Schooling Experience Family Income (1940 and 1950)	Mother's Schooling Father's Schooling Childhood I.Q. Family Income (1960)	6% 10%
		Schooling 1940, 1950 males	Mother's Schooling Father's Schooling Childhood I.Q. Siblings (in 1940)	Time Inputs in Early Childhood Siblings (1950) Family Income	8% 10%
		Schooling 1940, 1950 females	Mother's Schooling Father's Schooling Siblings Birth Order	Childhood I.Q. Family Income	17% 15%
		Childhood Ability I.Q. at c. age 11 males	Mother's Schooling Birth order Time inputs to child Family Income	Father's Schooling	19%
		Childhood Ability I.Q. at c. age 11 females	Family Income	Mother's Schooling Birth Order Time inputs to Child Father's Schooling	13%

Author	Sample	Dependent Variable	Related to	Not Related to	Variance Explained
Lindert	312 New Jersey executives 1008 observations	1963 Schooling age 31-81	Mother's Schooling Father's Schooling Father's Occupation Broken home Birth order No. of siblings Estimated Time Inputs	Age Sex	32% across siblings 8% between siblings
		1963 Occupational Status of 659 Males	Father's Schooling Father's Occupation Broken home Age Birth Order No. of siblings Estimated Time Inputs	Mother's Schooling	27%
Morgenstern	1968 Urban Problem Survey 842 blacks 782 whites men and women	1968 Average yearly Wage Rate	Schooling Experience Father's Schooling Sex School Quality for Blacks	School Quality for Whites	24% 27%
		Schooling 1968	Schooling Quality Region Mother's Schooling Father's Schooling Sex Age		21% 30%

Author	Sample	Dependent Variable	Related to	Not Related to	Variance Explained
Sewell and Hauser	Wisconsin Study 2069 men	1967 Earnings (age c.29) (from Socical Security)	Schooling Ability Occupation 1964 Parental Income (1957-1960)	Father's Schooling Mother's Schooling Father's Occupation	7%
		1964 Occupation measured by Duncan SEI	Schooling Ability Father's Occupation	Father's Schooling Mother's Schooling Parents' Income	41%
		Years of Schooling	Measured Ability Father's Schooling Mother's Schooling Father's Occupation Parents' Income		28%
		Measured Ability	Father's Schooling Mother's Schooling Father's Occupation Parents' Average Income		9%
Taubman	NAS-NRC	1973 Average Income of twins	Avg. Schooling Father's Schooling (MZ) Mother's Schooling (DZ) Hours Disability Status Self Employment Status Jewish Religion Catholic Religion (DZ twins only) Age (DZ)	Mother's Schooling (MZ) Father's Schooling (DZ) Age (MZ) Government Employee Other Religion	37% (MZ) 38% (DZ)

Author	Sample	Dependent Variable	Related to	Not Related to	Variance Explained
		1973 Income Difference between twins	Difference in Schooling Self Employment Status Government Employee (DZ) Disability Status (MZ) hours (if self employment, government status not included)	Government Employee (MZ) Disability Status	5% 12%
Wachtel	NBER-Thorndike 1812 men aged 44-52	1969 Income	Schooling Log of Hours Ability College Quality School Quality Experience Father's Education	Siblings Mother's Education	19%
		Schooling	Ability Siblings Father's Education Mother's Education Age	School Quality	10%
		Ability 1943	Mother's Education Father's Education School Quality Siblings		3%

Reviewing what is known, and deferring until later speculations on the unknown, the most consistent finding in the studies reviewed is that the variance in schooling accounts for a significant share of the explained variance in income. This is true for men in their 20's and 30's and was also true for different cohorts from ages 25 to 64. Labor market experience, when measured, never failed to be a significant predictor of earnings. Several studies found measured ability to be related to earnings (Duncan, Bowles and Nelson; Sewell and Hauser; Wachtel; de Wolff and Van Slijpe) but this was not the case in others (Conlisk; Leibowitz).

In three studies where several measures of family background were tried in earnings functions, family income proves to be the most significantly related to own income, while parental education and occupation do not retain significant coefficients if family income is in the equation. Sewell and Hauser found parents' income in 1957-60 the only significant background predictor of childrens' 1967 earnings. They also considered fathers' and mothers' schooling and fathers' occupation. Even income at the time of the child's birth was strongly related to child's income in a study by Conlisk. He found parental schooling levels not significantly related to income once parents' income 30 years prior was included in the equation. In Bowles and Nelson's study, parental income is a constructed variable, so it is not possible to separate its effect from father's schooling and occupation.

Both Conlisk and Sewell and Hauser use earnings data for ages 25-30, and this may cause the effect of family income to be overstated, since Leibowitz found that the effect of family income decreases with age, and became insignificant when the observations in her study were at an average age of 49. However, for the same group of men 10 and 20 years earlier, paternal income was significantly related to earnings, while other measures of background - mothers' and fathers' schooling - were not. In the studies where fathers' schooling or occupation is allowed to act as a proxy for earnings, it is generally a significant predictor of income.

What is the explanation for the consistent importance of family income in these studies? One unobserved variable for which paternal income may be a proxy is quality of schooling.

Wachtel found quality of high school and of college to be positively related to earnings. While Wachtel tried to test if his school quality variable was a proxy for family income, average per student expenditure in a school district may be a better index of the student's family income than is median income in the state because the variation in income across school districts within a state may be large compared to the variation within a school district. Per student school expenditure may also be a proxy for unmeasured family variables - such as the taste for schooling - since families can move from one school district to another in order to provide the school quality they demand.

Morgenstern found less direct measures of school quality not to be significantly related to earnings for whites, although they were for blacks. In discussing Sewell and Hauser's study above, it was suggested that family income might be inversely related with length of time required to complete a given level of schooling and would, therefore, be positively related to labor market experience. Alternatively, greater family income may allow financing of on the job training. Another variable, not considered in any of the studies reviewed, which has a positive impact on earnings and which is likely to be correlated with family income during childhood, is health. If, as some studies suggest (see Friedman and Leibowitz), families with greater income and education make greater investments in their children's health, which pays off in adulthood, we would expect own income to be related to parents' income, even holding other factors constant.

The studies using data on brothers imply a much larger role for the influence of family factors than has been demonstrated in the recursive models with direct measures of family background. This apparent paradox can be resolved by admitting that family variables matter - but that we don't have measures for all of them.

What might these family influences be? Taubman's paper points to factors such as disability status, self employment status, hours worked and religion which account for some of the variance in earnings. Differences in labor supply must be a large contributor to the variance in income, yet it is rarely explicitly accounted for. Choices about the amount of labor supplied to the market may prove to be a major mechanism whereby families influence the earnings of their children. If families are able to transmit "aggressiveness," "a taste for money income" or the "Protestant ethic" to their children, this may well show up as an increase in hours of work. Even without any correlation in wage rates, a correlation between fathers and sons in hours worked would generate the kind of relationship we have seen in several studies between parents' income and sons' income.

The relationship between parents' and children's hours of work might be elucidated by studying women's labor supply, which has greater variance than men's. The relationship between women's labor force participation and characteristics of their family of procreation has been amply examined, but the relationship with characteristics of the family of orientation has hardly been explored by economists. Using data on women we could, perhaps, differentiate the source of the income effect - attributing the impact of mother's earnings on her daughter's income to influence through the taste for hours of work and the impact of father's earnings to the quality of schooling or other income related factors.

Further work in this field will have to unmask the background factors which do prove to be determinants of economic success to determine <u>why</u> they matter. Lindert's tracing of the effect of family size on earnings to differential time inputs in childhood is an example of the kind of probing which is likely to be fruitful. Just as it would be hard to argue that employers are willing to pay workers for having had fewer siblings, it seems strained to argue that they would pay workers for having had richer families. The families' income must have been used to provide an attribute - such as higher quality schooling, or health - which increases productivity. These attributes cannot be generated costlessly, and thus the correlation of income between generations should not be considered as evidence of the existence of pure rents. In fact, if these attributes could be produced without cost, we would be less likely to find them correlated with income!

In spite of a great variety of independent variables measuring human capital, opportunity and ability, none of the studies reviewed above was able to account for much more than 50% of the variance in earnings or schooling. If we are unwilling to attribute more than half of the variance in earnings and schooling to chance, we will have to consider other sources of influence on measures of economic success.

The studies of brothers included in this volume employ a novel tactic. By dropping the assumption of a fully recursive model and allowing for common variance between brothers, these studies utilize more completely the information contained in samples of siblings. In this way, the importance of family background can be estimated without having knowledge of which particular facets of the family affect economic outcomes. These analyses amply demonstrate that there are one or more factors common to families which do affect economic success.

However, much work remains to be done in "naming" those factors. Naming with the labels "genetic" or "environmental," appears to be futile (as implicit in the papers of Chamberlain and Goldberger). But identifying the factors with processes in families so that we can understand the mechanism by which families influence their children, would be most worthwhile.

FOOTNOTES

[1] J. Mincer (1970), p. 2.

[2] Taubman and Wales did use four ability variables derived from a factor analysis of scores on 17 ability tests. The four factors correspond to mathematical ability, coordination, verbal ability and spatial perception. They found that only mathematical ability is a significant determinant of earnings (97-100). They note that "Thorndike believes our mathematical factor is close to I.Q.. . . . " (p. 97n).

[3] Exceptions include the paper by Chamberlain and Griliches discussed in Section II, and papers included in this volume.

[4] With the exception that father's occupation is second in importance in explaining son's occupation and of at least importance in explaining schooling and earnings.

[5] The single exception was that I.Q. was not a significant predictor of occupational status at ages 25-34 once years of schooling was included among the independent variables.

[6] Sewell and Hauser (1975), p. 79.

[7] The magnitude of this bias might be great since 1/3 of the students spent five years in obtaining a B.A. (p. 170) while the average number of years between college and obtaining the income data was six years.

[8] Sewell and Hauser (1975), p. 141.

[9] Sewell and Hauser find both standardized and regression coefficients of father's education exceed those of mother's education in predicting son's score on Henmon-Nelson Test of Mental Ability in his junior year in high school.

[10] The age range in the sample was 31 to 81, so most had completed their schooling.

[11] Chamberlain and Griliches (1975), p. 432.

REFERENCES

Becker, Gary S. Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education. New York. Nat. Bur. Econ. Res., 1964.

Becker, Gary S. "Human Capital and The Personal Distribution of Income." W. S. Woytinsky Lecture No. 1. Ann Arbor: University of Michigan, 1967.

Belmont, Lillian and Marolla, Francis. "Birth Order, Family Size and Intelligence," Science, 182, No. 4117 (Dec. 14, 1973).

Bowles, Samuel. "Schooling and Inequality from Generation to Generation," Journal of Political Economy, 80, No. 3, Pt. II (May/June 1972), S219-S251.

Bowles, Samuel and Nelson, Valerie I. "The Inheritance of I.Q. and the Intergenerational Reproduction of Economic Inequality." Rev. Econ. Stat. LVI, 1 (Feb. 1974), 39-51.

Chamberlain, Gary and Griliches, Zvi. "Unobservables with a Variance Components Structure: Ability, Schooling and the Economic Success of Brothers." International Econ. Rev. 16, 2 (June, 1975), 422-449.

Conlisk, John. "A Bit of Evidence on the Income-Education-Ability Interrelation." The J. of Human Resources, VI, 3 (Summer, 1971), 358-362.

Corcoran, Mary, Jencks, Christopher and Olneck, Michael. "The Effects of Family Background on Earnings. Am. Econ. Rev., 66, 2 (May, 1976), 430-435.

de Wolff, P. and Van Slijpe, A.R.D. "The Relation Between Income, Intelligence, Education and Social Background." European Econ. Rev. 4, (1973), 235-264.

Duncan, Otis Dudley. "Ability and Achievement." Eugenics Quarterly, 15 (March, 1968), 1-11.

Friedman, B. and Leibowitz, A., "Family Bequests and The Derived Demand for Health Inputs," mimeo, 1976.

Griliches, Zvi and Mason, William. "Education, Income and Ability," J. of Polit. Econ., 80, No. 3, Part II, p. S74-S103.

Hause, John. "Earnings Profile: Ability and Schooling," J. of Polit. Econ., 80 No. 3, Part II, p. S108-138.

Jensen, A. et. al., Environment, Heredity and Intelligence, Cambridge, Mass: Harvard Educational Review, 1969.

Leibowitz, Arleen. "Home Investments in Children." J. of Polit. Econ., 82, 2, Pt. II (March/April 1974), S111-S131.

Lindert, Peter. "Sibling Position and Achievement," University of Wisconsin, Madison, mimeo, 1976.

Mincer, Jacob. "The Distribution of Labor Incomes: A Survey with Special Reference to the Human Capital Approach." J. of Econ. Lit. 8, 1 (March, 1970) 1-26.

Mincer, Jacob. Schooling, Experience and Earnings. New York: NBER, 1974.

Morgenstern, Richard D. "Direct and Indirect Effects on Earnings of Schooling and Socio-Economic Background." Rev. Econ. Stat., LV, 2 (May, 1973), 225-233.

Sewell, William H. and Hauser, Robert M. Education, Occupation and Earnings: Achievement in the Early Career. New York: Academic Press, 1975.

Solmon, Lewis C. "The Definition of College Quality and its Impact on Earnings." Explorations in Economic Research, 2, No. 4 (Fall, 1975), 537-587.

Taubman, Paul. "The Determinants of Earnings: A Study of White Male Twins." University of Pennsylvania, mimeo.

Taubman, Paul and Wales, Terence. "Education as an Investment and a Screening Device." Education, Income and Human Behavior, F.T. Juster, ed. New York: Carnegie Commission and NBER, 1975.

Wachtel, Paul. "The Effect of School Quality on Achievement, Attainment Levels, and Lifetime Earnings." Explorations In Economic Research, 2, No. 4 (Fall, 1975), 502-536.

CHAPTER 3

CONTROLLING FOR AND MEASURING THE EFFECTS OF GENETICS AND FAMILY ENVIRONMENT IN EQUATIONS FOR SCHOOLING AND LABOR MARKET SUCCESS

J. Behrman[a], P. Taubman[a] and T. Wales[b]

Studies of the personal distribution of earnings have been addressed to a number of different questions.[1] These include: What are the sources of the inequality in earnings? Why is the earnings distribution skewed to the right? What will be the effects of small and large changes in particular variables such as schooling or strength on inequality? To what extent is the inequality attributable to innate abilities, to family environment, to imperfections in capital and other markets and to personally acquired skills? To what extent is the inequality due to health? Will policies directed towards improving equality of opportunity lead to sufficient equality of outcomes? And what is the relationship of annual and lifetime earnings?

The human capital model has been used to try to answer these questions. The basic idea in the human capital model is that real wages (net of any payments for investment in on-the-job training) equal marginal products. A person's marginal product depends on his skills and traits, which are produced by combining genetic endowments with environment. Particular aspects of the environment which can produce skills include, but are not limited to, family income, child rearing practices, number of siblings, school and on-the-job training. Often the human capital model is implemented by substituting the skill production function into the marginal product condition. This yields a relation between earnings and schooling and other variables.

Unfortunately, the samples available now and in the foreseeable future have little direct information on labor market skills (other than IQ) and very inadequate measures of genetic endowments and the types of environment listed above.

As is well known, the inability to control for genetic endowments and a variety of environmental variables can lead to biased estimates of the returns to education. This problem is often expressed as "the more educated are more able (irrespective of education) and the observed returns to education in part reflect the differences in ability." In this paper, we will use data on twins to control for genetics and family

---

[a]Professor of Economics, University of Pennsylvania

[b]Professor of Economics, University of British Columbia

We would also like to thank the Medical Follow-up Agency of the NAS-NRC for making the data available and conducting the survey and Drs. Norman, Hrubec, Mr. Simon and Miss Hanamura for their assistance. Professors Chamberlain, Goldberger, Griliches and Jencks have provided many suggestions and corrected numbers of errors. Barbara Atrostic and Shah Fardost supplied able research assistance.

environment. We will also estimate the bias from not controlling these variables.

As the title indicates, besides controlling for genetics and family environment, we are also interested in estimating the contribution of each of these variables to the variance of earnings and other variables. The question has been posed of why we are interested in such estimates.[2] Before giving the reasons that we believe important, we wish to dispose of several inappropriate ones. There is no question that transfers and taxes can make the distribution of income more equal even if genetics currently were to account for much of the inequality in pretax earnings. Some compensatory programs such as day care centers might help to equalize this distribution whether the sources of the existing inequality is genetics or family environment. Estimates of the current relative contribution of genetics in no way indicate what might be the relative contribution in the future, nor do they imply that environment can not be substituted for genetic endowments. Moreover, present genetically based inequalities are in no way socially natural. Society can, if it so chooses, use tax and transfer programs or compensatory training programs to redistribute income.

A basic question that society faces is whether the currently used tax, transfer, and compensatory training programs are optimal or whether changes in such redistributive policies are needed. A related question is whether policies aimed at improving the degree of equality of opportunity will lead to sufficient equality of outcomes. Such policy decisions are generally based on equity and efficiency criteria. We think that a portion of the equity criteria within and between generations is related to the extent to which a person's earnings are due to his own efforts versus those of the parents who bore and reared him.[3] Moreover nearly all the discussion of equality of opportunity concerns equalizing (or eliminating the worst parts of) family environment. It is of some interest to see how unequal the distribution would be if just family environments were equalized. This requires a separation of genetic and family environmental effects.

We suggest a tripartite breakdown of sources of inequality: genetics, common environment and specific environment. The sum of the first two categories are often labeled family effects. The techniques we use provide a reasonable estimate of the effects of specific environment and of family effects. As is common in many studies based on twins, we have more difficulty in separating the family effect into its genetic and common environment components. However, the techniques we use allow us to test certain key assumptions often made.

In this paper we will develop some methodologies that can be applied to samples of twins or, in some cases, a variety of relatives that will let us control for and measure the effects of genetics and family environment. We will also examine whether our results are robust under alternative model specifications. The structure of the model will allow us to determine whether education, occupation and earnings are related to the same or different genetic indices. We also try to contrast our model with an alternative one which in its most extreme form argues that genetic endowments have no effect on schooling or earnings.

I. Biological Aspects and Twins

By genetic endowments (G), we mean the innate capabilities that are based on a person's genes, half of which are contained in the egg and the other half in the sperm.[4] Environment (N) includes all other systematic and nonsystematic determinants of skills, including prenatal development. While environment is "everything else," some particular aspects that are usually thought to be important include family, peer group, on-the-job training, schooling and military. As this list indicates, an individual's environment includes both elements over which he has little and much choice.

Twin Types

Males and females normally have 23 pairs of chromosomes. The genes are located on the chromosomes with each parent contributing one gene of the pair found at each location. In the population each gene may come in one form such as A or in many varieties A, B...,Z. In an individual each of the two genes at a location may be the same or different, e.g., AA or AB. We will assume that each skill or trait is influenced by many genes, some of which have more than one variety. Only a randomly determined member of each gene pair is transmitted to the next generation via the egg or sperm, each of which is a gamete of one parent. But once the egg is fertilized, i.e., the two gametes combine, the developing individual receives one member of each gene from each gamete.

There are two types of twins-- monozygotic (MZ) and dizygotic (DZ). The MZs, often known as "identical," are the result of the splitting of an already fertilized egg, while the DZs, or "fraternal" twins, are the result of two different eggs fertilized by two different sperm. Thus, DZ pairs do not have the same genetic composition although they will be more alike than randomly drawn individuals. The MZ pairs, however, have the same genetic makeup because each piece of the split fertilized egg contains all and only the genetic information of the original fertilized egg (barring mutations).[5]

II. The Model

The general model we wish to estimate is given in Table 1. For simplicity in presentation we have assumed that there are no exogenous measured genetic or environmental variables. The model is easily modified to include such measures. The left-hand part of the table contains so-called structural equations, which may be thought of as solutions of supply and demand equations, while the right-hand side contains reduced-form equations.[6]

In these equations we represent the unobservables as G, G_1, G_2, G_3, N, N_1, N_2, and N_3, which we refer to as genetic and environmental indices.[7] An index can be written as $\Sigma_j b_j X_j$. The components (X_j) of each of the four genetic indices are the same, but their weights differ (and may be

Table 1

Model

	Structural Equations	Reduced Form
1)	$Y_1 = S = aG + bN + u$	$Y_1 = S = aG + bN + u$
2)	$Y_2 = OC_i + c'G + d'N + eG_1 + fN_1 + gS + u_1$	$Y_2 = OC_i = cG + dN + eG_1 + fN_1 + gu + u_1$
3)	$Y_3 = OC_{67} = h'G + j'N + k'G_1 + mG_2 + n'N_1 + pN_2 + qOC_1 + rS + u_2$	$Y_3 = OC_{67} = hG + jN + kG_1 + mG_2 + nN_1 + pN_2 + qu_1 + ru + u_2$
4)	$Y_4 = \ln Y_{73} = s'G + t'N + v'G_1 + w'G_2 + xG_3 + y'N_1 + z'N_2 + \alpha N_3 + z'N_2 + \alpha N_3 + \beta OC_{67} + \gamma S + u_3$	$Y_4 = \ln Y_{73} = sG + tN + vG_1 + wG_2 + xG_3$ $yN_1 + zN_2 + \alpha N_3 + \delta u + \lambda u_1 + \beta u_2 + u_3$

where

$c = c' + ag$

$d = d' + bg$

$h = h' + ra + qc$

$j = j' + rb + qd$

$k = k' + qe$

$n = n' + qf$

$s = s' + \beta h + a$

$t = t' + \beta j + \gamma b$

$v = v' + \beta k$

$w = w' + \beta m$

$y = y' + \beta n + \gamma b$

$z = z' + \beta p$

$\delta = \gamma + \beta r$

$\lambda = \beta q$

zero in some cases). It is easy to demonstrate that at most four separate indices are needed for our four equation model and that it is to a certain extent arbitrary which indices are included in which equation. The same statement holds for the environmental indices.

As noted earlier, in our earnings equation we assume that schooling, family and other environments and genetic endowments are combined to produce skills which determine a person's marginal product. Occupational status in 1967 is included in the equation, partly to capture the level of general and specific training given in various occupations, partly to capture trade-offs for occupational nonpecuniary costs and rewards and partly to capture windfall gains and losses for unexpected shifts in equilibrium wages for various occupations.[8] We assume that initial occupation only affects earnings via 1967 occupation since the latter occupation would incorporate the above effects and since we explicitly allow for G and N.

In some studies, occupational status is used as a proxy for annual or lifetime earnings. Since we have the status measures earlier in the earnings life cycle, our model is consistent with that view. We think, however, that a somewhat broader view in which people choose occupations to maximize their expected utility whose arguments include the wage and nonpecuniary payments from an occupation is appropriate.[9]

The years of schooling equation also incorporates supply and demand considerations. Becker has shown how family environment and innate abilities will affect the individual's demand for schooling. On the supply side, schools often select or reject students on the basis of ability.

The size and structure of this model is determined by a number of considerations. First, there is substantial interest in how individuals convert genetic endowments and family environment into labor market success. Second, there is a related interest in exploring how these genetic endowments and family environment change in their impact over the life cycle. Third, in the latent variable technique, which we discuss below, we need impose fewer restrictions the greater the number of observed covariances. Fourth, simplicity and manageability of handling data has led to limiting the current exploration to a few important indicators.

The earnings equation is semi log but the others are linear for two reasons. First, some tests for heteroskedasticity based on absolute values of differences for MZs indicate homoskedasticity for ln earnings but not for earnings. For the other three variables both the ln and linear versions display heteroskedasticity with variance increasing with the dependent variable in the linear version but decreasing in the ln version.[10] In principle it would be possible to find a transform such that the variances for Y_1, Y_2 and Y_3 are homoskedastic but the latent variable technique requires that the same transform of a particular variable appear everywhere in the model. Second, regression analysis indicated that semilog forms were better than double log forms for Y_1 through Y_4 but that linear forms were not grossly inadequate for the first three variables. Since much work in economics has used the semi log earning equations, we have kept it for comparability and used the linear forms for the other equations to aid in the latent variable technique.

In the next three sections, we indicate how we control G and N to obtain unbiased or consistent estimates of the parameters of the measured variables such as schooling in the earnings equation. We also indicate how we can estimate the contributions of the various G and N indices.

III. Controlling for Genetic Endowments and Family Environment

In previous studies of the effects of education on earnings, it has not been possible to control completely for the other determinants of earnings that might be correlated with schooling. With twins, however, it is possible to eliminate genetic differences for the identical twins and common background for both types by studying the within pair differences in earnings.

To understand what can and cannot be done with twins, it is necessary to compare the estimates obtained when using the individuals and within pair differences. As an aid in making this comparison, let us order individuals within each pair randomly, e.g., alphabetically, and denote the within pair difference by Δ . We can write an equation for individuals and a corresponding one within pairs as

1) $\ln Y = aS + bG + cN + u$

1a) $\Delta \ln Y = a\Delta S + b\Delta G + c\Delta N + \Delta u$

We can estimate both 1 and 1a and compare both OLS estimates of a. Denote the estimate from equation 1 as $\hat{a}_1$ and that from 1a as $\hat{a}_2$. Using standard methods, it can be shown that

2) $$\text{plim} \quad (\hat{a}_1) = a + \frac{\text{plim cov } (S, bG + cN + u)}{\text{plim var } (S)}$$

3) $$\text{plim} \quad (\hat{a}_2) = a + \frac{\text{plim cov } (\Delta S, b\Delta G + c\Delta N + \Delta u)}{\text{plim var } (\Delta S)}$$

As is well known, 2 yields biased estimates if plim cov (S, bG + cN + u) is nonzero, which is generally thought to be the case for the earnings, schooling model.

For MZ twins ΔG is zero. Making the usual assumption that Δu is uncorrelated with ΔS, plim $\hat{a}_2$ will be unbiased provided either c is zero or ΔN is uncorrelated with ΔS. The first condition means that the differences in MZ brothers' environments have no direct effect on earnings, though they may affect schooling. The second condition means that the differences in environment that determine earnings are not correlated with schooling. The latter condition may prevail if N consists of adult environment such as on-the-training. If the bias in $\hat{a}_1$ arises only because of genetics or common environment, $\hat{a}_2$ will be a con - sistent estimate.

For DZ pairs, ΔG is not zero and $\hat{a}_2$ will not be consistent if plim cov

(Δ S, bΔG) is not zero. Thus it is possible to determine if genetics should be controlled for by testing the null hypothesis that the MZ and DZ within equations are the same. It is interesting to note that the bias in $\hat{a}_2$ can be larger than that for $\hat{a}_1$ because both the numerator and denominator in the plim expression change when we go from levels to within pair differences.

In the introduction we indicated that some people argue that genetic endowments have no effect on earnings or schooling. As will be shown in the next section, they attribute the MZ-DZ difference in the sib correlation for Y to DZ's having a smaller correlation in environment than MZ brothers. In the above comparison of bias in the MZ and DZ within equations, we do not make use of the assumption that MZ and DZ brothers have the same environment correlation; thus, the test of the null hypothesis is a partial test of the hypothesis that there are no genetic effects.

This test is only partial in two respects. First, it is possible for genes to affect earnings but not bias the education coefficient in which case the MZ and DZ within equations would not differ. Second, our maintained hypothesis is that plim cov (ΔS, cΔN) is zero. If this is not true, it might explain why the two within equations were different. Fortunately, this latter possibility can be examined within the context of the latent variable models described below.

IV. Variance Components Model

In this section we will describe a variance components model that is often used to analyze twin data. We will conclude that if we examine only one variable such as ln earnings, the model is underidentified even if a number of strong and perhaps invalid assumptions are made. However, once we combine the variance component model with the latent variable technique, which will be introduced in the next section, we can identify the parameters and test the validity of most of these assumptions.

We will perform this analysis on an equation which can be expressed as

$$4)\quad Y = G + N$$

The reduced forms in Table 1 can be put into this format by combining the various genetic indices into one aggregate index (which can vary by dependent variable) and the N and the various error terms into a combined environmental index. We then normalize these indices so that their coefficients are each 1.

We can write the variance for individuals as

$$5)\quad \sigma_Y^2 = \sigma_G^2 + \sigma_N^2 + 2\sigma_{GN}$$

Now denote an MZ brother by an asterisk. We can then calculate cross sib covariances as

$$6)\quad \sigma_{YY*} = \sigma_{GG*} + \sigma_{NN*} + 2\sigma_{NG*}$$

Since for MZ twins G* = G, we can rewrite this as

$$7)\quad \sigma_{YY*} = \sigma_G^2 + \sigma_{NN*} + 2\sigma_{NG}$$

we can thus calculate that

$$8)\quad \sigma_Y^2 - \sigma_{YY*} = \sigma_N^2 - \sigma_{NN*} = (1-\rho*)\,\sigma_N^2 \text{ where}$$

$$\rho* = \sigma_{NN*}/\sigma_N^2$$

Thus the expected value of the difference between the individual and MZ cross sib covariance is the additive effect on Y of noncommon environment. Common environment presumably arises because of treatment in the family, neighborhood, school and elsewhere. Since the neighborhood and school are chosen by the family, it does not seem unreasonable to assume that the family influence is quite considerable in these common environmental effects. However, $\sigma_Y^2 - \sigma_{YY*}$ is an upper bound for nonfamily environment since parents may treat twins differently in relevant respects, an issue to which we will return shortly.

If we denote a DZ brother by a prime, we can also calculate

$$9)\quad \sigma_{YY'} = \sigma_{GG'} + \sigma_{NN'} + 2\sigma_{NG'}$$

Since for DZ twins, G does not equal G', other assumptions have to be made for us to be able to estimate other parameters. For example, $\sigma_{GG'} = {}^1/_2\sigma_G^2$ if all genetic effects are additive, if there is random mating, and if there are no sex linkages of genes. Also, $\sigma_{NN'} = \rho'\sigma_N^2$. Now it is assumed that $\rho* = \rho' = \rho$. Finally, assume that $N = \rho N' + v$ and E(G',v) = 0, so that $\sigma_{GN'} = \sigma_{GN}$. As we will stress in a moment, each of these assumptions are questionable, but even if they are true, the model still contains four parameters-- σ_G^2, σ_N^2, σ_{NG}, and ρ but only three observed statistics -- σ_Y^2, σ_{YY*}, and $\sigma_{YY'}$. Thus this model is underidentified.

Now let us consider these assumptions, beginning with those that yield $\sigma_{GG'} = {}^1/_2\sigma_G^2$. Various studies have indicated that on some traits, there is nonrandom or assortive mating of parents. As long as people choose to mate on the basis of the observed traits (the phenotype), and the observed traits partially reflect genetics, there will be correlation of the genes. Existing studies have shown positive assortive mating for measures of IQ and schooling and negative assortive mating for measures of personality such as extrovertism. Nonadditive genetic effects encompass both dominant, recessive genes and the effect of one gene depending on the level of another gene. Examples of both types of nonadditive effects exist in the genetics literature.[11]

The assumption that $\rho^* = \rho'$ has been questioned by a number of people. Evidence exists, as in Koch, that MZ twins are more likely to be dressed alike or treated alike in any one day than DZ twins. However, it has been argued that different observed parental treatment may indicate either that the parents of the DZ twins responded to the different genetically-based needs of the brothers or that it is because of their genes that the brothers selected different items from parental offerings.[12] Thus the observed differences in treatment are, in this view, due to genetic differences. Moreover, concentrating on the day-to-day treatment of twins or on dressing alike may be less appropriate than the year or childhood-long average treatment and quality of clothes. In addition, it is possible that the dimensions of family environment that are important are values and attitudes inculcated by the same examples to the brothers or good or bad nutrition offered jointly to both brothers. That is, the important aspects of environment may be like public goods which are not appropriable by one brother at the expense of the other.

However, some people still argue that σ_G^2 is zero and that $\sigma_{YY*} > \sigma_{YY'}$ only because $\rho^* > \rho'$. This view is difficult to accept since there are diseases that are caused by known genetic problems and result in mental retardation (if not treated) or are debilitating with respect to energy levels. Nevertheless, the general proposition that $\rho^* > \rho'$ may still be valid.

The assumption that $E(G',v) = 0$ can be interpreted as one DZ brother's noncommon environment being uncorrelated with his sib's genes. While this assumption may not be valid, it is not patently absurd. But note that this restriction has not been imposed on the MZ pairs.

We show in the next section that it is possible to estimate some of these coefficients in question rather than restricting them to particular values. We can then test these restrictions to determine their validity.

V. Latent Variables and Indicators

A "latent" variable is defined as an unobserved variable which affects two or more observed variables, called "indicators." While sociologists have studied such models for a number of years, econometricians' interest in this area dates from Zellner's pioneering study. Goldberger provides an extremely useful summary. Chamberlain and Griliches have used the technique with brothers, though they are not able to decompose "family effects" into genetic and environmental components.

In this section we will derive maximum likelihood estimators for a particular version of the model given in Table 1. Essentially we augment the standard latent variable model with a variance component model by including cross sib covariances. The reader not interested in the technical details may skip to the next section.

We refer to the right hand side of the four equations in Table 1 in which all the variables are unobservable as the reduced form residuals. Denoting the 4X1 vector of these reduced form residuals as ε and transposed by the superscript T, we define the following covariance matrices:

10) $E(\varepsilon\varepsilon^T) = \Omega_o \qquad E(\varepsilon\varepsilon'^T) = \Omega_D \qquad E(\varepsilon\varepsilon^{*T}) = \Omega_M$

where ε' is the residual vector for a DZ brother and ε^* is the residual vector for an MZ brother. The values of these matrices depend on the particular assumptions made about the covariances of the unobservables (N, N_1, N_2, N_3, G, G_1, G_2, G_3) and of the structural disturbances (u) as discussed below.

For estimation purposes we assume that the unit of observation is an individual and his twin brother, and that the corresponding 8X1 vector of reduced form residuals for the individual and his sib, $\binom{\varepsilon}{\varepsilon'}$ or $\binom{\varepsilon}{\varepsilon^*}$ is denoted as e' or e* for DZ and MZs respectively.[13] Similarly, we define the corresponding covariance matrices Ω' and Ω^* as:

11) $E(e'e'^T) = \Omega' = \begin{pmatrix} \Omega_o & \Omega_D \\ \Omega_D & \Omega_o \end{pmatrix}$

12) $E(e^*e^{*T}) = \Omega^* = \begin{pmatrix} \Omega_o & \Omega_D \\ \Omega_D & \Omega_o \end{pmatrix}$

Assuming now that e' and e* are independently normaly distributed, then the logarithm of the likelihood of observing a sample with (T' + 1) and (T* + 1) DZ and MZ twins respectively is given by (aside from a constant)

13) $L = -\frac{T'}{2}(\log/\Omega'/ + tr[\Omega'^{-1} W']) - \frac{T^*}{2}(\log/\Omega/ + tr[\Omega^{*-1} W^*])$

where W' and W* are defined analogously to Ω' and Ω^*, except that their elements are the observed sample covariances of the four dependent variables.[14] This likelihood function can be maximized with respect to the underlying structural parameters in the reduced form parameters in Table 1 once Ω' and Ω^* (or, more basically, Ω_o, Ω_D and Ω_M) are expressed in terms of these parameters and provided the parameters are in fact identified.

Before discussing this table, we will consider the question of identification of the model.[15] A subset of our model can be expressed as

14) $Y_1 = \lambda_1 G + \gamma_1 N + u_1$

$Y_2 = B_1 Y_1 + \lambda_2 G + \gamma_2 N + u_2$

$Y_3 = B_2 Y_1 + \alpha Y_2 + \lambda_3 G + \gamma_3 N + u_3$

$$Y_4 = B_3Y_1 + \delta Y_3 + \lambda_4 G + \gamma_4 N + u_4$$

For this discussion it is useful to decompose G and N into the family and individual specific components, $G_{ij} = f_i + g_{ij}$, $N_{ij} = m_i + h_{ij}$, where i is the family and j is the sib subscript. We will assume that $E(f_i m_i) = \sigma_{fm}$ and $E(g_{ij}h_{ij}) = \sigma_{gh}$ while the specific components are uncorrelated with the family components. We will also assume that the u are uncorrelated across equations, across brothers and with G and N.

The reduced form of this model is

$$15)\quad (Y_1 Y_2 Y_3 Y_4) = (\underset{\sim}{d}\ \underset{\sim}{k}) \begin{bmatrix} G \\ N \end{bmatrix} + (E_1, E_2, E_3, E_4)$$

with

$$\underset{\sim}{d} = \begin{bmatrix} \lambda_1 \\ \lambda_2 + B_1\lambda_1 \\ \lambda_3 + \alpha(\lambda_2 + B_1\lambda_1) + B_2\lambda_1 \\ \lambda_4 + \delta[\lambda_3 + \alpha(\lambda_2 + B_1\lambda_1) + B_2\lambda_1] + B_3\lambda_1 \end{bmatrix}$$

$$\underset{\sim}{k} = \begin{bmatrix} \gamma_1 \\ \gamma_2 + B_1\gamma_1 \\ \gamma_3 + \alpha(\gamma_2 + B_1\gamma_1) + B_2\gamma_1 \\ \gamma_4 + \delta[\gamma_3 + \alpha(\gamma_2 + B_1\gamma_1) + B_2\gamma_1] + B_3\lambda_1 \end{bmatrix}$$

$$\underset{\sim}{E} = \begin{bmatrix} E_1 \\ E_2 \\ E_3 \\ E_4 \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 + B_1u_1 \\ u_3 + \alpha(u_2 + B_1u_1) + B_2u_1 \\ u_4 + \delta(u_3 + B_2u_1 + \alpha(u_2 + B_2u_1)) + B_3u_1 \end{bmatrix}$$

and $E(\underset{\sim}{E}\underset{\sim}{E}') = V =$

$$
\begin{array}{cccc}
\sigma_1^2 & B_1\sigma_1^2 & (B_2 + \alpha B_1)\sigma_1^2 & (\delta\alpha B_1 + B_2\delta + B_3)\sigma_1^2 \\
 & \sigma_2^2 + B_1^2\sigma_2^2 & B_1(B_2 + \alpha B_1)\sigma_1^2 + \alpha\sigma_2^2 & (\delta(\alpha B_1+B_2) + B_3)B_1\sigma_1^2 + \alpha\delta\sigma_2^2 \\
 & & \sigma_3^2 + (B_2 + \alpha B_1)^2\sigma_1^2 + \alpha^2\sigma_2^2 & \delta\sigma_3^2 + \alpha^2\delta\sigma_2^2 + (\alpha B_1 + B_2)(\delta(\alpha B_1 + B_2) + B_3)\sigma_1^2 \\
 & & & \sigma_4^2 + \delta^2\sigma_3^2 + \alpha^2\delta^2\sigma_2^2 + (\alpha\delta B_1 + \delta B_2 + B_3)^2\sigma_1^2
\end{array}
$$

and where $\sigma_1^2 = \sigma_{u_1}^2$, etc.

Thus we can express our model in terms of the individual and cross sib variance covariance matrices whose probability limits are given by

16) $\Omega_o = (\sigma_f^2 + \sigma_g^2)\underset{\sim}{d}\underset{\sim}{d}^T + (\sigma_m^2 + \sigma_h^2)\underset{\sim}{k}\underset{\sim}{k}^T + (\sigma_{fm} + \sigma_{gh})(\underset{\sim}{d}\underset{\sim}{k}^T + \underset{\sim}{k}\underset{\sim}{d}^T) + V$

17) $\Omega_D = \sigma_f^2 \underset{\sim}{d}\underset{\sim}{d}^T + \sigma_m^2 \underset{\sim}{k}\underset{\sim}{k}^T + \sigma_{fm}(\underset{\sim}{d}\underset{\sim}{k}^T + \underset{\sim}{k}\underset{\sim}{d}^T)$

18) $\Omega_M = (\sigma_f^2 + \sigma_g^2)\underset{\sim}{d}\underset{\sim}{d}^T + \sigma_m^2 \underset{\sim}{k}\underset{\sim}{k}^T + (\sigma_{fm} + \sigma_{gh})(\underset{\sim}{d}\underset{\sim}{k}^T + \underset{\sim}{k}\underset{\sim}{d}^T)$

At times it is convenient to use the within sib covariances:

19) $\Sigma_M = \Omega_o - \Omega_M = \sigma_h^2 \underset{\sim}{k}\underset{\sim}{k}^T + V$

20) $\Sigma_D = \Omega_o - \Omega_D = \sigma_g^2 \underset{\sim}{d}\underset{\sim}{d}^T + \sigma_h^2 \underset{\sim}{k}\underset{\sim}{k}^T + V$

Now let us consider how we can estimate the variance parameters. Note first that we can only estimate $\sigma_1^2, \sigma_2^2, \sigma_3^3$, and σ_4^2 and B_1, B_2, B_3, α and δ from V since the first four parameters appear nowhere else, while there is not enough information to identify the others elsewhere. Since V contains 10 independent elements, it can be used to estimate these nine parameters. It is worth noting that V only contains 9 parameters because we assume that initial occupation (Y_2) does not enter into the structural equation for earnings (Y_4). We do allow initial occupation to affect 1967 occupation and the latter enters the earnings equation.

Now consider the Σ_M expression. Since it contains V plus $\sigma_h^2 \underset{\sim}{k}\underset{\sim}{k}^T$, we can estimate the 9 elements in V plus σ_h^2 (with k estimated in Ω_M or Ω_D).[16]

In practice we estimate the items in V and σ_h^2 from both Σ_M and Σ_D thereby increasing efficiency. But to use both these expressions, we must partial $\sigma_g \underset{\sim}{d}\underset{\sim}{d}$ out of Σ_D. The potential information to estimate σ_g^2 and the rest of the parameters in the model is contained in Ω_D and Ω_M.

Following Chamberlain it can be shown that since Ω_D has rank 2 it has 7 independent variables. Ω_M is also of rank 2 with 7 independent elements. But again following Chamberlain it can be shown that only 3 of these 7 elements are independent of those in Ω_D. Ω_M and Ω_D contain the remaining 13 unknown parameters, σ_f^2, σ_m^2, σ_{fm}, σ_{gh}, λ_1, λ_2, λ_3, λ_4, γ_1, γ_2, γ_3, γ_4, and B_4. If σ_m^2 varies for MZs and DZs, we will have 14 parameters to estimate.

Thus with the 10 independent covariances in Ω_M and Ω_D and the two normalizations we are allowed, we have 12 elements to estimate the 14 parameters given above. In practice we resolve the underidentification by restricting four of the parameters. That is, we start off with a model in which we assume all covariances of unobserved variables are zero, $\sigma_{GN}^2 = 0$ for both MZs and DZs and the genotypic correlation for DZs = 1/2. We then test to see if dropping any one of these restrictions improves the fit of our model or alters the estimates of the other parameters.

The above comment applies to a model with one G and one N indicator. If we allow more such indices, it can be shown that there will be enough information to estimate up to 20 parameters outside of the Σ_M, Σ_D segment, in which 10 additional parameters can be estimated.

We have been discussing identification in a model with one G and one N index. Suppose we were to add a second genetic index which appeared in the equations for the last three indicators. Both Ω_D and Ω_M will have rank 3. Therefore, taken separately each will have 9 independent elements. However, it can be shown that only 6 elements in Ω_M would be independent of those in Ω_D. Thus the blocks will have 15 independent elements while the blocks will still have 10. If we normalize $\sigma_{G_1}^2$ to be one, we can estimate its coefficient in the equations for the last 3 indicators and any of the 2 parameters such as σ_{GG_1}, σ_{GN} or the equivalent of σ_f^2/σ_G^2. We cannot, however, reduce the underidentification in the 1G, 1N index part of the model. Similarly, we can add G_2 and G_3 to the model as in Table 1 and estimate additional parameters.

Next, suppose we return to the 1G, 1N model and then add N_1 to our model in the form given in Table 1. Once again the ranks of Ω_D and Ω_M change so that they will contain 15 independent variables. Suppose we decompose N into family and specific components, $m_{1_i} + h_{1_{ij}}$. Then it can be shown that $\Sigma_M = \sigma_h^2 \underset{\sim}{k}\underset{\sim}{k}^T + \sigma_{h_1}^2 \underset{\sim}{k_1}\underset{\sim}{k_1}^T + V$. (The elements in k_1 can be found in Table 1.) As previously indicated, Σ_M contains only 10 independent elements and V has 9 unknown parameters. Thus we can only estimate one of the 4 remaining unknowns, σ_h^2, k, $\sigma_{h_1}^2$, k_1. Unfortunately, σ_h^2 and $\sigma_{h_1}^2$ do not appear in Ω_D or Ω_M. While k and k_1 appear in the Ω blocks, they always multiply σ_m^2 and $\sigma_{m_1}^2$, respectively. Thus it is not possible to identify the coefficients unless further restrictions such as $\sigma_{h_1}^2 =$ 0 or $\sigma_h^2/\sigma_N^2 = \sigma_{h_1}^2/\sigma_{N_1}^2$ are introduced.

As noted earlier, we can incorporate the full set of genetic effects in our 4 indicator model by having a single vector that contains all the possible genetic elements and then constructing 4 indices out of this vector by using various weights. To estimate the parameters in V and σ_h^2, we must partial out all the equivalents of σ_G^2 for each of the 4 indicies from Σ_D, which can be estimated from the Ω_D and Ω_M blocks.

It is equally true that 4 N indices would be needed to cover fully all environmental effects. But since we can only estimate one of σ_{hj}^2, j = 1. . .4, we will concentrate on models with only one environmental index. Note, however, that the estimates of the genetic effects are derived from $\Omega_D - \Omega_M$. Even if there are 4 environmental indices, this differencing will eliminate their contributions, provided the cross sib correlation is the same for MZs and DZs for each environmental index.

Finally, this model can be compared with one in which we first partial out the effect of a set of observable exogeneous variables. That is, assume

$$21)\qquad N = X_N \mu_N + v_1$$

$$22)\qquad G_p = X_{G_p} \mu_{G_p} + v_2 \qquad p=1...4$$

where the Xs are observables that may serve as <u>proxies</u> for N and the Gs, the s are unknown coefficients, and the v's are random disturbances.

Then the method presented above can be followed with the v's replacing N and the Gs. To the extent that the Xs are good proxies for the unobservables, the variances of the v's will be small relative to those of N.

VI. The Sample

In this study we will use the NAS-NRC twin sample, which is described in more detail in Appendix B, which also contains the questionnaires. Briefly, however, for this study we have a maximum of 2,478 pairs where each brother answered a questionnaire mailed in 1974.[18] Most of these pairs also answered several earlier questionnaires to which we have access. To be included in the mailing, the twins had to be born between 1917 and 1927, be white, have served at some point in the military and be alive in 1974. These last restrictions suggest an under-representation of people with low intelligence or education, or with poor mental and physical health as compared to the corresponding white male cohort. Even compared to a population of veterans, the respondents to our questionnaire have more earnings and education.[19,20] Regression estimates from stratified samples with nonpopulation weights still yield unbiased coefficients over the sample space.[22]

In some of our analysis, we assume that MZ and pairs in our sample are random drawings from the same population. The DZ pairs may have a different distribution of genes and/or environment because DZ pairs as a percentage of births occur more frequently among older women. The corresponding percentage for MZ pairs is independent of mother's age and SES class.[23]

The means and variances of earnings, schooling and several other variables by twin types are given in Table 2. It is evident that in our sample, the DZ twins come from families in which the parents have a bit less education and occupational status, and in which the number of siblings and older siblings are somewhat greater, and some of these differences are statistically significant. The religious distributions are also very similar for MZs and DZs, which is a bit surprising since we expected Catholics to have more children at older ages and thus to be a larger portion of the DZ pairs. The means of schooling, initial and later occupational status and earnings are nearly the same across twin type although the variances differ by up to 10%.[24]

The conclusion that MZ and DZs are random drawings from the same population is further strengthened by a comparison of the simple correlations given in Table 3.[25] The left-hand portion of that table treats both brothers as individuals and the results are close for MZs and DZs. The right-hand portion consists of cross sib correlations, defined for example as $\sigma_{SY'}/\sigma_S \sigma_Y$ where $\sigma_{SY'}$ is the covariance of one brother's years of schooling and his sib's ln of earnings. The cross sib correlations are uniformly lower than the comparable ones for individuals. The DZ cross sib correlations are uniformly lower than the comparable MZ ones.

VII. Ordinary Least Squares Regressions

We have computed ordinary least squares regressions for individuals

Table 2

Some Summary Statistics for Individuals in the NAS-NRC Sample (calculated separately for MZ's and DZ's)

	MZ's		DZ's	
	Mean	Variance	Mean	Variance
1973 annual earnings	18.4[a]	150[b]	18.1[a]	166[b]
ln 1973 annual earnings	9.67[a]	.28	9.64[a]	.32
1967 or 1972 occupational score[d]	50.4	472	49.8	445
Years of schooling	13.5	9.1	13.3	9.8
Initial full time civiliam occupation[c,d]	36.7	610	35.0	590
Age	51.0	8.4	51.2	8.8
Mother's education years	10.0	9.6	9.7	11.9
Father's education years	9.3	12.6	9.1	14.8
Father's occupational status[d]	29.6	532	28.6	503
% Catholic	26	19	23	18
% Jewish	4	4	5	5
% Other non-Protestant	2	2	3	3
Number of siblings alive 1940	2.6	4.9	3.0	5.6
Number of older siblings alive 1940	1.6	3.3	2.1	3.7
Number of pairs	1019		907	

Note. Calculations are for those for whom earnings are non-zero for both brothers. For other variables, if one brother answered and the other did not, non-respondent is set equal to his brother. If both did not answer, both are set at mean or put in "other category". For mother and father data, if brothers' answers differ, mean of responses is used.

[a]Thousands of $

[b]Millions of $

[c]As recalled in 1974.

[d]Duncan Scale

Table 3

Individual and Cross-Sib Correlations

		Individuals				Cross-Sib			
		S	OC_i	OC_{67}	lnY	S	OC_i	OC_{67}	lnY
MZ's									
	S	1	.528	.542	.442	.764	.467	.440	.406
	OC_i		1	.450	.349		.525	.351	.322
	OC_{67}			1	.350			.429	.268
	ln Y				1				.545
DZ's									
	S	1	.530	.507	.441	.545	.365	.295	.292
	OC_i		1	.436	.360		.333	.220	.225
	OC_{67}			1	.351			.205	.187
	ln Y				1				.295

Note: Correlations for individuals calculated as $\frac{\sigma_{xy}}{\sigma_x \sigma_y}$

Correlations for sibs calculated as $\frac{\sigma_{xy'}}{\sigma_x \sigma_y}$

Note: Calculated for both brothers having nonzero earnings 1973. For other variables, missing observations are replaced with brother's values, or if both are missing, with the mean. Less than 20 people did not report years of schooling. About 500 people did not report initial occupation. About 800 people did not report 1967 occupation. A comparison of cross-sib correlations for OC_{67} based only on people who reported indicated slightly smaller figures for MZs and DZs with the difference between the two cross sib correlations practically unchanged.

S is years of schooling.

OC_i is initial civilian occupation, scored on the Duncan Scale.

OC_{67} is occupation in 1967, scored on the Duncan Scale.

ln Y is the natural log of earnings in 1973.

and within MZ and DZ pairs for the various dependent variables. In the equations based on individuals we have also estimated some equations which contain a number of proxies for genetic endowments and family environment and one so-called reduced form which contains only these proxies.

We will consider first the question of the bias that arises if we do not control at all for common environment and genetics or partially control with the available proxies.[26] Subsequently we will consider the implications of the various proxies used. The equations for individuals are given in Table 4 and those for MZ and DZ within pairs in Table 4a.

For individuals the coefficient on years of schooling is about .08 whether or not age is included. Since our sample is not a random drawing of the population it is interesting to note that studies based on Census data yield similar estimates.[27] When we included OC_{67} we divide up the total effect of schooling into a direct effect of .07 and an indirect effect of about .01. In the DZ within equations in which we control for common environment, the estimate of the total effect of education declines to .06. In the MZ within equation, in which we control for common environment and genetic endowments, the total effect of education declines to less than .03. We have used an analysis of covariance (Chow test) to test the null hypothesis that the MZ and DZ within equations are the same. For both the equation with and without OC_{67}, this null hypothesis is rejected at the 5% level.[28]

We have also estimated an equation for individuals in which we have included a number of variables which often are used as proxies for family environment and genetic endowments. While most of these proxies have significantly nonzero coefficient estimates, the total effect of education only declines to .07. The implied bias of 2/3 between the individual and MZ within equation probably is an upper bound because the variance of measurement error to variance in true years of schooling is almost surely greater in the within equations. But as indicated in Appendix A, the bias will remain for likely magnitudes of measurement error.[29]

In the first equation for OC_{67} for individuals, the coefficient on schooling is about .36. Once initial occupation is included the direct effect of schooling is .29. In the MZ and DZ within equations, the education coefficients are reduced 20 and 10%, respectively. The Chow test indicates that the within equations are not significantly different at the 10% level. The inclusion of the proxy variables in the individual equation does not change the education coefficient.

For the initial occupation, the coefficient on years of schooling is about .42 or slightly greater than in the OC_{67} equation. The coefficient drops to .28 in the DZ within equation and .21 in the MZ equation. These two equations do not differ significantly. Once we include our proxy variables in the individual equations, the coefficient of education declines from .42 to .37.

Next let us consider the proxy variables. In the reduced form equations, most of these variables are significant and have the same sign for Ed, OC_i, OC_{67} and $\ln Y_{73}$. Most of these findings are in accord with

Table 4
Structural and Reduced Form Equations Estimated for Individuals by OLS
From runs 2/10/76

Yrs. of Schooling	S	OC_i	OC_{67}	Age	Raised Rural	Married 1974	Cath	Jew	Born South	# Sibs alive 1940	ED_F	ED_M	OC_F	Con-stant	$\bar{R}^2$
(S-1) S				-.05 (3.5)	-.41 (4.3)	.23 (1.9)	-.24 (2.3)	1.61 (7.8)	-.01 (.1)	-.20 (10.2)	.12 (8.1)	.10 (6.4)	.017 (9.4)	13.82 (17.7)	.19
Initial Occupation															
(OCI-1) OC_i	.42 (40.8)													-2.01 (14.7)	.28
(OCI-2) OC_1				.0092 (.7)	-.31 (4.0)	.08 (.8)	.11 (1.3)	1.22 (7.2)	.46 (4.8)	-.14 (8.7)	.086 (7.0)	.066 (4.9)	.0095 (6.3)	1.66 (2.6)	.12
(OCI-3) OC_i	.37 (33.2)			.028 (2.6)	-.15 (2.2)	-.0030 (.03)	.20 (2.6)	.62 (4.1)	.46 (5.4)	-.064 (4.4)	.041 (3.6)	.027 (2.2)	.0030 (2.2)	-3.49 (5.9)	.30
1967 Occupation															
(OC67-1) OC_{67}	.36 (40.0)													.134 (1.1)	.27
(OC67-2) OC_{67}	.29 (27.2)	.19 (14.6)												.50 (9.0)	.31
(OC67-3) OC_{67}				-.021 (1.9)	-.61 (8.8)	.15 (1.8)	-.10 (1.3)	.77 (5.2)	.30 (3.6)	-.075 (5.4)	.040 (3.5)	.043 (3.2)	.011 (8.1)	5.16 (9.0)	.10
(OCI-4) OC_{67}	.28 (25.0)	.18 (13.7)		-.006 (.7)	-.45 (7.4)	.077 (1.1)	-.030 (.5)	.23 (1.8)	.23 (3.2)	.004 (.3)	.0086 (.8)	-.010 (.9)	-.004 (3.4)	1.10 (2.1)	.32
lnY73															
(Y1) lnY_{73}	.080 (32.4)													8.58 (262.5)	.20
(Y2) lnY_{73}	.067 (23.6)		.039 (9.5)											8.55 (262.8)	.22
(Y3) lnY_{73}				-.011 (4.1)	-.054 (3.0)	.14 (6.6)	.032 (1.6)	.43 (11.2)	.007 (.32)	-.021 (6.0)	.014 (5.2)	.012 (4.0)	.0018 (5.3)	9.8 (67.5)	.11
(Y4) lnY_{73}	.059 (19.5)		.035 (8.8)	-.0079 (3.1)	-.008 (.5)	.13 (6.3)	.05 (2.6)	.29 (8.3)	-.0080 (.4)	-.0082 (2.4)	.0082 (3.0)	.0050 (1.6)	.0004 (1.3)	8.84 (63.5)	.25

Note: OC_{67} and OC_i divided by 10 and are scaled from 0 to 10 as compared to Table 2.

Table 4a
Structural Equations
Within Pair, MZ and DZ Separately, OLS

	MZ							DZ						
	ΔS	ΔOC_i	ΔOC_{67}	Married$_1$	Married$_2$	Constant	R^2	ΔS	ΔOC_i	ΔOC_{67}	Married	Married	Constant	R^2
OC_i														
(OCI-1)	2.10			-.126	.047	.061	.03	.28			-.15	.24	.10	.09
	(5.9)			(.6)	(.2)	(.3)		(9.5)			(.6)	(1.0)	(.4)	
(OCI-2)	.209					.13	.03	.28					-.023	.09
	(5.9)					(82.5)		(9.5)					(31.3)	
OC_{67}														
(OC67-1)	.26	.15		-.15	.12	.013	.11	.29	.14		-.14	-.42	.43	.19
	(8.4)	(5.8)		(.9)	(.7)	(.08)		(11.3)	(5.5)		(.7)	(2.2)	(1.9)	
(OC67-2)	.26	.15				-.009	.11	.29	.14				-.04	.19
	(8.4)	(5.8)				(2.7)		(11.3)	(5.4)				(48.3)	
(OC67-3)	.29					-.027	.08	.33					-.05	.16
	(9.5)					(33.4)		(13.4)					(259.5)	
lnY73														
(Y-1)	.017		.026	-.10	-.13	-.02	.04	.048		.038	.063	-.10	.035	.09
	(2.2)		(5.1)	(3.1)	(2.7)	(.4)		(6.0)		(3.8)	(1.0)	(1.9)	(.5)	
(Y-2)	.019		.040			.0005	.03	.048		.038			-.002	.08
	(2.4)		(4.9)			(.7)		(6.0)		(3.9)			(4.8)	
(Y-3)	.026			.124	-.095	-.020	.02	.059			.052	-.12	.057	.07
	(3.5)			(2.9)	(2.2)	(.4)		(8.3)			(.9)	(2.0)	(.8)	
(Y-4)	.027					.0030	.01	.059					-.00094	.07
	(3.6)					(9.2)		(8.2)					(5.2)	

Note: These regressions use moments calculated from a single entry for each pair, i.e., 1 minus brother 2. Intercepts are not forced to zero as would be done if double entry methods were used, i.e. brother 1 minus brother 2 as one observation and brother 2 minus brother 1 as another observation.

OC_{67} and OC_i divided by 10 as compared to Table 2.

Variable Definitions

S	is years of schooling, reported in 1974
OC_i	is initial full-time civilian occupation, reported in 1974, scaled on the Duncan Score
OC_{67}	is current occupation, reported mostly in 1967 but later for some of the sample, scaled on the Duncan Score
$\ln Y_{73}$	is the natural log of annual earnings in 1973, reported in 1974
Age	is 1974 minus birth date, taken from birth certificates
Raised Rural	is a dummy variable equal to 1 if raised in rural districts, reported in 1967
Married 1974	is a dummy variable equal to 1 if married in 1974, reported in 1974
Catholic	is a dummy variable equal to 1 if raised in Catholic religion, reported in 1974
Jewish	is a dummy variable equal to 1 if raised in Jewish religion, reported in 1974
Born South	is a dummy variable equal to 1 if born in the Census defined region of the South, taken from birth certificates
# Sibs Alive in 1940	is number of sibs alive in 1940, reported in 1974
ED_F	is years of schooling of father, reported in 1974
ED_M	is years of schooling of mother, reported in 1974
OC_F	is father's occupation, Duncan Score, reported in 1967

previous findings, e.g., parental education and occupation have positive effects. The number of siglings alive in 1940 has negative effects for all four variables, although not significantly different from zero for education. Those who are Jewish receive more education and achieve more occupational status and earnings. The Catholic variable has a significantly negative effect on education. The rural variable has uniform negative effects. In the education equation, age is negative, presumably because of the differential opportunities available to different cohorts. The negative effect on earnings may represent a cohort effect but is consistent with the peaking of age earnings profiles around age 45 found m cross-sections. (Age is negative in earnings equations even when S and OC_{67} are included.) The marital status variable probably is a proxy for current family income needs and significant only in the earnings equation.

Once education and occupation are included as independent variables, the coefficients on these proxies are often reduced, sometimes changed in sign and become less significant. Thus, much of the effects of family background takes place through or are mediated by schooling and occupation.

VIII. Latent Variable Results—Tests of Some Models

In this section we examine a number of alternative versions of the latent variable model. We are forced to examine all these models because we do not have enough information to estimate the most general model that could be specified. If we find that relaxing a particular assumption does not significantly change the value of the likelihood function and does not affect greatly parameters of interest, we will impose the restriction in subsequent runs. Because many of the runs are very similar and because of the costliness of obtaining standard errors, in most runs we did not calculate t statistics for structural parameters or for constrained parameters in the reduced forms. The reader not interested in the detailed comparisons may turn to the summary of this section that commences on page 68.

As is explained in the section on identification, the subsystem that contains all the observed variables and one G and one N index would have in its most general form 24 parameters to be estimated from 20 independent observations and 2 normalizations. To estimate the model and provide a basis for testing restrictions, we initially impose the following additional restrictions: $\sigma_{NG} = \sigma_{NG'} = \sigma^2_h = 0$ and $\sigma_{G_iG_i'} = 1/2\ \sigma^2_{G_i}$, i = 0, 1, 2, 3 (i.e., random mating and all gene effects are additive). Also, we convert all the cross sib covariances into correlation coefficients with $\sigma^2_m/\sigma^2_N =$ and $\sigma^2_h/\sigma^2_N = (1-\rho)$. Finally for reasons discussed earlier, we included three additional orthogonal genetic indices. The resulting model has 23 unknown parameters, once we normalize $\sigma^2_N = \sigma^2_G = \sigma^2_{G_1} = \sigma^2_{G_2} = \sigma^2_{G_3} = 1$, to be estimated from 30 covariances.

The estimates for this model are given in Table 5. While we do not

Table 5
Four Indicator Model (23 Parameters)

	G	N	G_1	G_2	G_3	u	u_1	u_2	u_3	S	OC_i	OC_{67}	$\ln Y_{73}$
Reduced Form Equations													
S	2.090 (20.5)	1.710 (13.7)				1							
OC_i	.847 (7.0)	1.043 (8.2)	1.143 (17.1)			.197 (5.7)	1						
OC_{67}	.814 (7.3)	.658 (5.4)	.361 (4.8)	.794 (13.0)		.290 (5.7)	.149 (8.8)	1					
$\ln Y_{73}$	.210 (7.8)	.150 (5.1)	.102 (5.6)	.016 (.6)	.300 (23.1)	.024 (3.1)	.0045 (3.6)	.030 (3.2)	1				
Structural Equations													
S	2.090 (20.5)	1.710 (13.7)				1							
OC_i	.436 (2.7)	.706 (5.7)	1.143 (17.1)				1			.197 (5.7)			
OC_{67}	.144 (.9)	.058 (.5)	.192 (2.0)	.794 (13.0)				1		.261 (5.7)	.149 (8.8)		
$\ln Y_{73}$	.155 (4.2)	.105 (3.8)	.091 (5.0)	-.0078 (.3)	.300 (23.1)				1	.015 (1.9)		.030 (3.2)	

Restrictions and other Parameter estimates

Normalizations and Restrictions

$\sigma^2_G = \sigma^2_{G_1} = \sigma^2_{G_2} = \sigma^2_{G_3} = \sigma^2_N = 1;$

$\sigma^2_{N_1} = \sigma^2_{N_2} = \sigma^2_{N_3} = 0$

$\rho_{MZ} \quad \rho_{DZ} = 1,$

$\sigma_{NG} = \sigma_{NG'} = \sigma_{G_1N} = \sigma_{GG_1} = \sigma_{G_1N'} = 0,$

AM = 1/2

Other Estimates

$\sigma^2_u = 2.18$ (22.6)

$\sigma^2_{u_1} = 2.80$ (25.2)

$\sigma^2_{u_2} = 2.49$ (26.2)

$\sigma^2_{u_3} = .13$ (24.0)

Functional Value = 13435.44

wish to comment in detail on the results for each version of our model, we will examine our first effort with some care. The latent variable technique requires that the same unmeasured variable be significant in at least two of the reduced form equations. The t statistics, given in parentheses, on all the G and N terms are significant at the 5% level and often at the 1% level except for G_2 in the equation for lnY.[30] Thus it appears there is one common genetic index and one common environmental index that affects all four variables, directly or indirectly. There is also an orthogonal genetic index that affects OC_i and lnY_{73} and two others, one of which affects only OC_{67} and the other of which affects lnY_{73}.

The structural equations allow us to separate out the direct and indirect effects of each variable. Once again, nearly all the coefficients on the genetic and environmental indices are significant.

In the various structural equations all of the coefficients on the measured variables have the expected sign and are significant at the 5% level. If besides G only common environment has direct effects on Y or the OC variables, then the MZ within equations will yield consistent estimates of the parameters on the measured variables in our recursive system. These MZ within equations, which are given in Table 4a, yield estimates which are quite comparable, differing by no more than .02. As noted earlier, the MZ within estimates are very different from those obtained from the equations for individuals.

IX. Variations on a Theme

The previous results are based on a particular restricted version of our model. Next we will relax some of these restrictions.

To help keep the notation clear let us redefine $\rho^* = \rho_{MZ}$ and $\rho' = \rho_{DZ}$. We begin by letting $\rho_{MZ} = \rho_{DZ} \neq 1$. As shown in Table 5a, the maximum likelihood estimate of is .87, which has a standard error of .11. Thus ρ is 8 standard deviations from zero but 1.3 standard deviations from 1. The ln of the likelihood function has changed by .3.[31] Using the likelihood ratio test with 1 degree of freedom, twice this difference is not significant at the 5% level. Thus we do not reject the hypothesis that no part of noncommon environment has direct effects on all 4 indicators.

In comparison with the estimates when is restricted to 1, the new coefficients on the unobserved variables are unchanged for the G indices and increased by about 8% for the N index. When ρ is less than 1, the model should be interpreted as saying that some environment that is specific to one brother affects all the variables. Thus it is not surprising that the coefficients on N are bigger since σ_{τ}^2 is still normalized at 1.[32] More importantly under this interpretation, MZ within equations estimated one by one no longer need be consistent since differencing across brothers will no longer eliminate all omitted variables correlated with the independent variables. Thus it is encouraging to find that the coefficients on S in the structural equations only decrease by small numerical amounts, though this is enough to reduce the direct effect of schooling to zero in the lnY equation.

Table 5a
Four Indicator Model
(24 Parameters)

	G	N	G_1	G_2	G_3	u	u_1	u_2	u_3	S	OC_i	OC_{67}	$\ln Y_{73}$
Reduced Form Equations													
S	2.092 (20.5)	1.831 (10.9)				1							
OC_i	.843 (7.0)	1.121 (7.7)	1.140 (16.9)			.097 (.6)	1						
OC_{67}	.814 (6.0)	.704 (5.0)	.361 (4.7)	.793 (13.3)		.267 (5.8)	.145 (5.1)	1					
$\ln Y_{73}$	.208 (7.7)	.163 (5.1)	.093 (4.4)	.020 (.7)	.302 (23.5)	.0079 (.3)	.0043 (3.0)	.029	1				
Structural Equations													
S	2.092	1.831				1							
OC_i	.640	.943	1.140				1			.097			
OC_{67}	.164	.080	.196	.793				1		.253	.145		
$\ln Y_{73}$	.184	.142	.083	-.0032	.302				1	.0010		.029	

Normalization and Restrictions

$\sigma^2_G = \sigma^2_{G_1} = \sigma^2_{G_2} = \sigma^2_{G_3} = \sigma^2_N = 1, \quad \sigma^2_{N_1} = \sigma^2_{N_2} = \sigma^2_{N_3} = 0$

$\sigma_{NG} = \sigma_{NG'} = \sigma_{GG_1} = \sigma_{G_1N} = \sigma_{G_1N'} = 0,$

AM = 1/2

Other Estimates

$\rho_{MZ} = \rho_{DZ} = .86$ (8.2)

$\sigma^2_u = 1.75$ (4.0) $\qquad \sigma^2_{u_2} = 2.49$ (26.1)

$\sigma^2_{u_1} = 2.71$ (17.5) $\qquad \sigma^2_{u_3} = .13$ (21.5)

Functional value = 13435.14

As the model is structured, we cannot identify separate estimates for ρ_{MZ} and ρ_{DZ}. (See Goldberger's paper for a formal proof.) We can, however, restrict ρ_{DZ} to be a fixed fraction of ρ_{DZ}. The results obtained when $\rho_{DZ} = .8 \rho_{MZ}$ are given in Table 5b. The estimates of the variances of the random errors (u through u_3) and the coefficients of the measured variables are identical in Tables 5a and 5b. In 5b the estimates of the coefficients on G and G_1 are smaller while those on N are larger when $\rho_{MZ} > \rho_{DZ}$. All the coefficients significant in Table 5a remain so in 5b except for that on G in the equation for OC_i. The fit of the model is, of course, the same as in 5a and is not a significant improvement over that in Table 5.

Till now we have been assuming that $\sigma_{GN} = \sigma_{GN'} = 0$, which restriction we would like to drop. Unfortunately, we have not been able to get this version of the model to converge, apparently because the gradient of the likelihood function is very flat with respect to σ_{GN}. It is possible, however, to estimate the coefficients on σ^2_N for any assigned value of σ_{GN}. The difference in the likelihood function when $\sigma_{GN} = 0$ or .6 is only .05. The coefficients on N vary greatly between the runs in which $\sigma_{GN} = 0$ and .6. Examples of these differences are given in Tables 5d and 5f in which we have also relaxed the composite restriction of random mating and only additive gene effects. The reader will note that only when $\sigma_{GN} = .6$ the coefficients on σ^2_N change, a result which can be shown analytically to be valid. We also have not been able to estimate σ_{GN} and $\sigma_{GN'}$ when ρ is not 1.

We have replaced the random mating additive gene effects assumption by one that says for each genetic index for DZ pairs that $\sigma_{GG'}/\sigma^2_G = \sigma_{G_1G_1'}/\sigma^2_{G_1} = \sigma_{G_2G_2'}/\sigma^2_{G_2} = \sigma_{G_3G_3'}/\sigma^2_{G_3} = AM$ where AM is estimated. Tables 5d, 5e and 5f give the results for this model for ρ restricted to 1 or allowed to vary and for various values of σ_{GN}.

The introduction of the AM parameter causes a sharp change of about 3.5 for the ln of the likelihood function. The estimate of AM is about .35 which implies negative assortive mating or alternatively nonadditive genetic effects. When approached in terms of assortive mating, the value of .35 is somewhat surprising since for IQ and schooling, positive assortive mating has been found. Of course, for personality traits such as extroversion, negative assortive mating has been found.[33] The results are more explicable in terms of dominance which reduces DZ cross sib genetic correlation below 1/2—assuming random mating.[34] As should be expected, compared with the random mating model, the estimates of genetic effects are decreased while those for N are increased. But the estimates for the parameters obtained primarily from Σ_M and Σ_D are

Table 5b
Four Indicator Model
(24 Parameters)
$\rho_{DZ} = .8\,\rho_{MZ}$

	G	N	G_1	G_2	G_3	u	u_1	u_2	u_3	S	OC_i	OC_{67}
Reduced Forms												
S	1.56 (7.2)	2.30 (12.2)				1						
OC_i	.37 (1.4)	1.41 (8.0)	1.07 (7.7)			.097 (.7)	1					
OC_{67}	.61 (2.6)	.89 (5.3)	.38 (2.8)	.78 (8.8)		.27 (5.4)	.15 (5.4)	1				
$\ln Y_{73}$	.17 (3.0)	.21 (5.1)	.11 (2.7)	.015 (.3)	.30 (14.2)	.0080 (.4)	.0043 (3.3)	.029	1			
Structural Equations												
S	1.56	2.30				1						
OC_i	.22	1.19	1.07				1			.097		
OC_{67}	.176	.10	.23	.78				1		.25	.15	
$\ln Y_{73}$	.15	.18	.10	-.0079	.30				1	.00009		.029

Restrictions

$\sigma^2_{N_1} = \sigma^2_{N_2} = \sigma^2_{N_3} = 0$ $\qquad$ AM = 1/2

$\sigma^2_G = \sigma^2_{G_1} = \sigma^2_{G_2} = \sigma^2_{G_3} = \sigma^2_N = 1$ $\qquad$ $\sigma_{NG} = \sigma_{NG'} = 0$ $\qquad$ $\rho_{DZ} = .8\rho_{MZ}$

Other Estimates

$\rho_{MZ} = .91$ (1/.8) $\qquad$ $\sigma^2_u = 1.75$ (3.8) $\qquad$ $\sigma^2_{u_2} = 2.49$ (26.2)

$\sigma^2_{u_1} = 2.71$ (17.4) $\qquad$ $\sigma^2_{u_3} = .129$ (21.6)

Function Value 13435.14

Table 5c
Four Indicator Model
(24 Parameters)
AM ≠ 1/2

	G	N	G_1	G_2	G_3	u	u_1	u_2	u_3	S	OC_i	OC_{67}
Reduced Forms												
S	1.85 (15.9)	1.98 (17.5)				1						
OC_i	.68 (5.5)	1.16 (11.7)	1.16 (18.5)			.21 (6.0)	1					
OC_{67}	.69 (6.3)	.78 (8.4)	.37 (5.1)	.82 (13.8)		.29 (8.9)	.14 (5.4)	1				
lnY	.17 (5.9)	.19 (7.7)	.098 (6.0)	.019 (.8)	.31 (26.2)	.026 (3.4)	.0044 (3.5)	.031 (4.7)	1			
Structural Equations												
S	1.85 (15.9)	1.98 (17.5)				1						
OC_i	.30 (1.9)	.75 (5.7)	1.16 (18.6)				1			.21 (6.0)		
OC_{67}	.11 (.1)	.09 (1.0)	.20 (2.3)	.82 (13.8)				1		.26 (7.9)	.14 (5.4)	
lnY	.12 (3.3)	.13 (5.2)	.087 (5.2)	-.0068 (.3)	.31 (26.1)				1	.016 (2.3)		.031 (4.7)

Restrictions and Other Parameter Estimates

Restrictions

$\sigma_G^2 = \sigma_{G_1}^2 = \sigma_{G_2}^2 = \sigma_{G_3}^2 = \sigma_N^2 = 1$

$\rho_{MZ} = \rho_{DZ} = 1, \quad \sigma_{NG} = \sigma_{NG'} = 0$

$\sigma_{N_1}^2 = \sigma_{N_2}^2 = \sigma_{N_3}^2 = 0$

Other Estimates

AM = .34 (6.1)

σ_u^2 = 2.17 (22.6) $\quad \sigma_{u_1}^2$ = 2.75 (24.4)

$\sigma_{u_2}^2$ = 2.45 (25.1) $\quad \sigma_{u_3}^2$ = .127 (23.1)

Functional Value = 13431.87

Table 5d
Four Indicators
(25 Parameters)
AM ≠ 1/2
ρ ≠ 1

	G	N	G_1	G_2	G_3	u	u_1	u_2	u_3	S	OC_i	OC_{67}
Reduced Forms												
S	1.85 (15.9)	2.07 (13.2)				1						
OC_i	.67 (5.5)	1.22 (10.3)	1.16 (18.3)			.12 (.9)	1					
OC_{67}	.70 (6.3)	.82 (7.8)	.37 (5.1)	.82 (13.9)		.27 (5.2)	.14 (5.3)	1				
lnY	.17 (5.9)	.20 (7.3)	.09 (4.6)	.022 (.9)	.31 (26.3)	.0096 (.4)	.0042 (3.2)	.030	1			
Structural Equations												
S	1.85	2.07				1						
OC_i	.45	.97	1.16				1			.12		
OC_{67}	.13	.12	.20	.82				1		.28	.14	
lnY	.15	.17	.079	-.002	.31				1	.0016		.030

Restrictions

$$\sigma^2_G = \sigma^2_{G_1} = \sigma^2_{G_2} = \sigma^2_{G_3} = \sigma^2_N = 1$$

$$\sigma_{NG} = \sigma_{NG'} = 0$$

$$\sigma^2_{N_1} = \sigma^2_{N_2} = \sigma^2_{N_3} = ($$

Estimates

AM = .34 (6.1)

$\rho_{MZ} = \rho_{DZ} = .90$ (10.3)

$\sigma^2_u = 1.76$ (3.8) $\quad \sigma^2_{u_1} = 2.67$ (18.3)

$\sigma^2_{u_2} = 2.45$ (25.0) $\quad \sigma^2_{u_3} = .12$ (20.0)

Functional Value = 13431.61

Table 5e
Four Indicator
(26 Parameters)
BM, AM, $\rho_{MZ} = \rho_{DZ} \neq 1$

	G	N	G_1	G_2	G_3	u	u_1	u_2	u_3	S	OC_i	OC_{67}
Reduced Forms												
S	1.811	2.103				1						
	(13.2)	(12.6)										
OC_i	.650	1.230	1.163			.123	1					
	(4.9)	(10.4)	(18.4)			(.974)						
OC_{67}	.681	.828	.368	.824		.271	.140	1				
	(5.9)	(7.9)	(5.1)	(13.8)		(5.0)	(5.3)					
lnY	.168	.202	.0908	.021	.310	.009	.0042	.030	1			
	(5.9)	(7.6)	(4.6)	(.92)	(25.7)	(.4)	(3.1)					
Structural Equations												
S	1.81	2.10				1						
OC_i	.550	.971	1.163				1			.123		
OC_{67}	.130	.122	.205	.824				1		.254	.140	
lnY	.147	.174	.080	-.0031	.310				1	.0020		.030

Restrictions

$$\sigma^2_{N_1} = \sigma^2_{N_2} = \sigma^2_{N_3} = 0$$

$$\sigma^2_G = \sigma^2_{G_1} = \sigma^2_{G_2} = \sigma^2_{G_3} = \sigma^2_N = 1$$

$$\sigma_{NG} = \sigma_{NG'} = 0$$

Estimates

$\rho_{MZ} = \rho_{DZ} = .90$ (10.3)

AM = .29 (3.6) BM = .38 (4.9)

$\sigma^2_u = 1.77$ (3.7) $\sigma^2_{u_1} = 2.67$ (18.2)

$\sigma^2_{u_2} = 2.44$ (24.8) $\sigma^2_{u_3} = .125$ (19.6)

Functional value = 13431.52

Table 5f
Four Indicator
(24 Parameters)
AM, σ_{GN} = .6

	G	N	G_1	G_2	G_3	u	u_1	u_2	u_3	S	OC_i	OC_{67}
Reduced Forms												
S	1.853 (15.6)	1.148 (8.6)				1						
OC_i	.681 (5.4)	.780 (6.4)	1.184 (21.7)			.205 (5.99)	1					
OC_{67}	.695 (5.9)	.466 (4.0)	.362 (5.5)	.822 (14.2)		.294 (5.9)	.144 (8.9)	1				
lnY	.172 (5.9)	.113 (3.8)	.097 (6.6)	.020 (.9)	.309 (26.9)	.026 (3.4)	.0045 (3.5)	.0308	1			
Structural Equations												
S	1.850	1.148				1						
OC_i	.301	.563	1.184				1			.205		
OC_{67}	.597	.0471	.1909	.822				1		.264	.144	
lnY	.150	.079	.0856	.071	.309				1	.0198		.0308

Restrictions

$$\sigma^2_{N_1} = \sigma^2_{N_2} = \sigma^2_{N_3} = 0$$

$$\sigma^2_G = \sigma^2_{G_1} = \sigma^2_{G_2} = \sigma^2_{G_3} = \sigma^2_N = 1$$

$$\rho_{MZ} = \rho_{DZ} = 1$$

$$\sigma_{GN} = \sigma_{GN'} = .6$$

Estimates

AM = .34 (6.2)

σ^2_u = 2.17 (24.6) $\qquad \sigma^2_{u_1}$ = 2.74 (24.5)

$\sigma^2_{u_2}$ = 1.24 (25.1) $\qquad \sigma^2_{u_3}$ = .127 (23.0)

Functional value = 13431.93

little changed.

We have also estimated a model with AM for the first three genetic indices but BM for the last genetic index. The results given in Table 5e are not much different from those with AM restricted to be the same for all four genetic indices.

Up to now we have estimated our models with 4 G and 1 N indices. While in principle it is necessary to have as many genetic indices as indicators,[35] in practice it may not be necessary to be so generous. Similarly, it is possible to add more environmental indices provided we restrict the ρ on new indices to be 1 (or the same as on the first N index). In a table not shown, restricting all the coefficients on G_1 to be zero reduces the ln of the likelihood function by about 34 compared to the model in Table 5. Since twice this difference is highly significant, we can fit the data much better with G_1 included. As might be expected, the coefficients on the other parameters estimated primarily in the Ω_D, Ω_M blocks shift about when G_1 is excluded from the model. The coefficients derived primarily from the Σ block tend to increase by small amounts.

Next we return to the model with four genetic indices and add a new orthogonal index (N_2) to the OC_{67} and Y_{73} equations in which its ρ is restricted to 1.[36] The new variable has coefficients with t values of 1.93 and .5. However, the coefficients on N_2 are numerically small, the other coefficients are essentially unchanged, and twice the ln of the likelihood function changes by about .8, which is not significant at the 5% level.

Up to now we have started with models which include several genetic indices. Since some people take seriously a model in which only environment matters, we estimate a system in which ρ_{MZ} is not restricted to ρ_{DZ} for the various N indicators. Following Chamberlain, this system can be shown to be identified if it has no more than 21 parameters. Moreover, the earlier material on identification indicates that in such a model we can only estimate one ρ_{MZ} and one ρ_{DZ}.

Table 5g contains a model in which for the four N indices we constrain ρ_{MZ} to be the same and similarly ρ_{DZ} is constrained to be the same, through $_{MZ}$ need not equal ρ_{DZ}. Here we find that the estimates of the coefficients on the observables and for σ^2_u, $\sigma^2_{u_1}$, $\sigma^2_{u_2}$ and $\sigma^2_{u_3}$ are quite close to those obtained in earlier runs in which $\rho_{MZ} = \rho_{DZ} \neq 1$ (Tables 5a and 5d), though several of the coefficients are no longer significant. The log of the likelihood function is smaller than our previous best estimates by 16. Moreover, our estimates are $\rho_{MZ} = .95$ and $\rho_{DZ} = .61$.

Once we introduce a single genetic index (assuming AM = 1/2) into this pure environmental model, the log of the likelihood function returns to approximately the same level as in our earlier "genetic" runs. Thus one

Table 5g
4 N Indices
(21 Parameters)

	N	N_1	N_2	N_3	u	u_1	u_2	u_3	S	OC_i	OC_{67}
<u>Reduced Forms</u>											
S	2.76				1						
	(31.5)										
OC_i	1.36	1.17			.095	1					
	(23.2)	(18.8)			(.9)						
OC_{67}	1.07	.34	.79		.26	.15	1				
	(20.8)	(5.2)	(12.3)		(5.4)	(5.3)					
lnY	.26	.09	.026	.30	.0053	.0042	.029	1			
	(20.1	(5.0)	(11.2)	(19.3)	(.2)	(3.1)					
<u>Structural</u>											
S	2.76				1						
OC_i	1.10	1.17				1			.095		
OC_{67}	.18	.17	.79				1		.25	.15	
lnY	.24	.08	.003	.30					-.0021		.029

<u>Restrictions and Other Estimates</u>

$\sigma^2_G = \sigma^2_{G_1} = \sigma^2_{G_2} = \sigma^2_{G_3} = 0$ $\quad$ $\sigma_{NG} = \sigma_{NG'} = 0$

$\sigma^2_u = 1.75$ (4.2) $\quad$ $\sigma^2_{u_1} = 2.73$ (16.11) $\quad$ $\sigma^2_{u_2} = 2.51$ (24.3)

$\sigma^2_{u_3} = .129$ (13.5) $\quad$ $\rho_{MZ} = .94$ (14.9) $\quad$ $\rho_{DZ} = .61$ (18. 8)

Functional value = 13447.27

can explain the data significantly better with genetics included. The estimates for ρ_{MZ} and ρ_{DZ} change little. The new coefficients and those on N are insignificant.

X. Summary of Variations

We began the latent variable section with a model which embodies the same assumptions as those made in estimating the MZ within equations. We then studied a number of variations which involve changes in restrictions in either or both of the Σ or Ω blocks. The coefficients on the observed variables and the σ_u^2 come primarily from the Σ blocks while the others come primarily from the Ω_M, Ω_D blocks. In this summary, we will concentrate on the effects of these changes on the various parameters and on whether relaxing a restriction significantly improves the fit of the model.

For the estimates of the parameters contained in the Σ blocks, the initial model we fitted should yield results very similar to those obtained from the MZ within since in the Σ blocks this model essentially augments the MZ within information by the corresponding but adjusted DZ information. (The adjustment consists of the partialling out of the DZ genetic effects which are estimated in the Ω_D or Ω_M block.) The results for the estimates of the parameters on the observed variables and the σ_u^2 are indeed very close. As long as we retain four genetic indices, the relaxing of the other restrictions in the Ω_D or Ω_M blocks only changes our estimates of the DZ genetic effects marginally. (Thus, relaxing these restrictions has little effect on the estimates obtained from the Σ blocks.)

Put another way, varying assumptions about genotypic correlation for DZ's and correlation between N and G yield essentially the same results for comparable parameters as those obtained in the MZ within equations, which are much simpler and cheaper to estimate. The extra efficiency obtained from combining MZ and DZ material, of course, may be very valuable when the total sample or particular cells are small. If the only objectives in using twin data are to control for common environment and genetic endowments and to estimate (an upper bound to) noncommon environment, MZ within equations or the latent variable model with any one of the set of assumptions we studied can be used.

The latent variable model can be used to examine the proposition that the environment that affects all (several) of our indicators includes part of each sib's specific environment. This is an important modification because differencing across MZ brothers does not eliminate specific environments, whose presence might cause a bias in one or more of the structural equations. The argument that only common environment appears in all the structural equations implies that $\rho = 1$. We have tested this restriction in several variants of our model. We find point estimates of ρ to be around .85 to .9. As would be expected, when ρ is less than 1, the parameters on schooling are reduced but the estimated coefficients on the genetic terms are little changed. While the point estimates for ρ are .85 to .9, these coefficients are not significantly different from 1. Also compared to estimates in which $\rho=1$, twice the difference in the log of the likelihood function is less than 1.0, which is not very significant (using a Chi square test with 1 degreee of freedom). Not wishing to be accused of trying to drive the rate of return on schooling to zero, we will not reject the hypothesis that the MZ within coefficients are unbiased-- subject to a subsequent discussion on measurement error.

The estimates of the parameters obtained primarily from the Ω block are more sensitive to the relaxation of the other restrictions. The dropping of the restriction that the DZ coefficients on the genetic indices are 1/2 decreases the coefficients on the first genetic index by about 20 to 30% with comparable increases in the N coefficients.[37] Since twice the difference in the log of the likelihood function is about 7, it is not difficult to decide to reject the model with the coefficients restricted to 1/2.

A second restriction is that $\sigma_{NG} = \sigma_{NG'} = 0$. When $\rho = 1$ we can maintain the assumption that $\sigma_{NG} = \sigma_{NG'}$, but as we indicated earlier we have not been able to get our estimates to converge. For different values of σ_{GN}, the coefficients of N vary greatly but the remaining coefficients are not changed.

Finally, we have tried to estimate models in which ρ_{MZ} is not restricted to ρ_{DZ} but we can not identify the two ρ's in this model. When we restrict all genetic effects to zero, and include four N indices but restrict each N to have the same ρ_{MZ} and ρ_{DZ}, our estimates for ρ_{MZ} and ρ_{DZ} are about .95 and .61, respectively. While this model yields estimates of the coefficients of schooling, etc., quite close to those in the MZ within equations, the model does not fit the data as well as the mixed genetic/environment models.

In the next section we will examine in more detail the results in several of our models and consider their implications.

XI. Implications, Caveats and Conclusions

Our results have implications for several questions of interest, though there are some important caveats which we discuss below. The first question is the importance of education in achieving occupational status and earnings. In discussing this topic we will confine ourselves to total effects which are obtained by substituting the appropriate equations in our recursive system.

Consider first the effect of schooling on earnings which is about .025 once we adjust for G and N.[38] Under certain conditions, which need not be met in this sample (nor for that matter, in any other sample), this coefficient is the rate of return to schooling. This estimate is, of course, much lower than the before tax returns available to individuals on comparable investments. However, before jumping to the conclusion that education is a bad investment, we must add that the conditions required for the .025 to be the private return to schooling include that all returns to schooling accrue in the form of increased wages. If some returns occur as nonpecuniary job rewards or outside the maket place, then the .025 may be a very understated estimate of the rate of return to education.

As noted earlier the .025 estimate is substantially less than the .08 obtained when we do not control for G and N. Since we show that generally available proxies for family background are inadequate as controls, there is a possibility that prior studies have yielded seriously biased estimates of the returns to schooling.

Educational attainment also helps explain initial and 1967 occupational status. The education coefficient is about 3/4 as large for initial occupation as in 1967, while the mean status level is about 2/5 as large initially. For reasons which are not clear to us, omitting G and N causes a larger bias for initial occupation than for occupation in 1967.[39]

Both the earnings and occupational status results on the importance of earnings and on bias from omitting G and N may be affected by measurement error but, as shown in Appendix A, such error will probably not affect the tenor of our results.

Next let us consider the contributions of our unmeasured variables. Since the results are somewhat sensitive to the model specification, we will present several alternatives. Table 6 is based on the model in which we allow for assortive mating and nonadditive genetic effects but in which we restrict $\sigma_{NG} = \sigma_{NG'} = 0$ and $\rho_{MZ} = \rho_{DZ} = 1$. In this version, genetics accounts for about 30% of the variance in occupational status and 45% of the variance in 1973 earnings. Common environment (N) accounts for 42% of the variance in schooling, 22% in initial occupation and 10% in 1967 occupation and 1973 earnings. The remainders, which are attributable to specific environment, account for 25 to 55% of the total.

Of the genetic effects, most of the total (direct plus indirect) impacts are attributable to the index introduced in an equation for a particular indicator, e.g., G_2 for OC_{67}. However, the first genetic index has noticeable effects on all four indicators while the second accounts for 3% of the variance in OC_{67} and in Y_{73}. The information in Table 5d indicates, moreover, that the genetic indices have significant direct effects in the structural equations.

When σ_{GN} is not restricted to zero, the coefficients on N vary greatly but the others are unchanged. The results are shown in Table 6a. An allocation of the σ_{GN} term between G and N effects will be arbitrary, but any such allocation would, if anything, increase the genetic contribution.

The estimates of the genetic contribution are affected by our assumptions about assortive mating and dominance. As can be inferred from Table 5, estimates of the effects of the first G index would be about 25% greater if we assumed that there was random mating and additive genetic effects.

The estimates would not change greatly if we were to use the

Table 6

Sources of Variance of Schooling, Initial and Later Occupational Status and Earnings

(Assuming $\sigma_{NG} = \sigma_{NG'} = 0$)

	Schooling	Initial Occupation	1967 Occupation	1973 Earnings
σ	36%	8%	11%	10%
$\sigma^2_{G_1}$		23	3	3
$\sigma^2_{G_2}$			15	0.01
$\sigma^2_{G_3}$				32
$\Sigma\sigma^2_G$	36	31	29	45
σ^2_N	41	22	13	12
σ^2_u	23	02	04	0.5
$\sigma^2_{u_1}$		46	01	0.02
$\sigma^2_{u_2}$			53	0.8
$\sigma^2_{u_3}$				42

Source table 5c AM = .35, $\rho_{MZ} = \rho_{DZ} = 1$

Table 6a

Sources of Variances of Schooling, Initial and Later Occupations and Earnings

(Assuming $\sigma_{NG} = \sigma_{NG'} = .6$)

	Schooling	Initial Occupation	1967 Occupation	1973 Earnings
σ^2_G	36%	08%	11%	10%
$\sigma^2_{G_1}$		23	3	3
$\sigma^2_{G_2}$			15	0.01
$\sigma^2_{G_3}$				32
$\Sigma\sigma^2_G$	36	31	29	45
σ_{NG}	26	11	07	08
σ^2_N	14	11	05	04
σ^2_u	23	02	04	0.04
$\sigma^2_{u_1}$		46	01	0.01
$\sigma^2_{u_2}$			55	0.40
$\sigma^2_{u_3}$				42

Note: $\rho_{MZ} = \rho_{DZ} = 1$

Source Table 5f AM = .35

$\sigma_{GN} = \sigma_{GN'} = .6$

variants in which we had only three genetic indices or in which we add a second environmental index with a ρ of 1. If in our basic model we do not restrict ρ to 1, we transfer part of the specific environmental effect from the errors uncorrelated over equations to the N term which is correlated over equations. Other than such a change, this model's results would be unaltered from above.

When ρ_{DZ} is resticted to .8 ρ_{MZ}, the contributions of genetics decrease. As shown in Table 6b, genetics would still contribute importantly to all four dependent variables.

In the pure environmental model, the estimates of family effects would be equal to the sum of the G and N effects in Table 6.

In Table 4 we have some equations which include a number of commonly available proxies for G and N. The R^2 in the reduced form equations are far less than the sum of the genetic and common environment effects obtained from the latent variable method.

Perhaps the best way to summarize this section is that all our various models (except the pure environmental one) suggest that genetic endowments, common environment and specific environment all contribute to schooling and labor market success. Common environment appears more important for schooling and early occupation, while genetics is more important for earnings and schooling. If these results can be taken as even approximately correct, they imply that even extreme policies to assure equality of opportunity by eliminating all differences in common environment (including those due to the family) would not eliminate much of the family contribution to the welfare of offspring-assuming we know how to eliminate all differences in "common environment".

These results, of course, are obtained from a particular sample of white male veterans born between 1917 and 1927. This sample has a different mean and variance in earnings and education than either the population as a whole or the white men of the same age group. This noncomparability may mean that our results, especially those for the partition of the variance, do not apply to the cohort or for the population as a whole. Moreover, even if they do apply to the current population, there is no reason to consider any of these effects fixed.

As we indicated in the introduction, we can redistribute income through taxes and transfers or through compensatory training programs. The question is , should we redistribute and how should we go about the redistribution. The answer to this question is complicated because most schemes that improve the equity of the income distribution involve losses in efficiency.

Our results do not indicate the effects of various policies of efficiency, but they do give some information on part of what we consider to be equity. All our models indicate that the total of genetic and family environment account for more than half of the total variance in earnings around age 50.[40] Assuming that combined family effects are this large for lifetime income for current generations and that the standard error in the ln of earnings was about .55, we would conclude that a substantial amount of redistribution was required to obtain equality. Such redistribution can in principle be accomplished by compensatory programs,

Table 6b

Sources of Variances of Schooling, Initial and Later Occupations and Earnings

(Assuming $\rho_{DZ} = .8\ \rho_{MZ}$)

Percent of Total Arising from	S	OC_i	OC_{67}	$\ln Y_{73}$
σ^2_G	25%	02%	08%	09%
$\sigma^2_{G_1}$		19	03	04
$\sigma^2_{G_2}$			14	0.07
$\sigma^2_{G_2}$				30
σ^2_G	25	21	25	43
σ^2_N	56	33	17	14
σ^2_u	18	0.2	03	0.04
$\sigma^2_{u_1}$		45	01	0.02
$\sigma^2_{u_2}$			54	0.7
$\sigma^2_{u_3}$				43

Source Table 5b

elimination of capital market imperfections, or by taxes and transfers. The evidence in this paper suggests more schooling won't be very helpful in equalizing the income redistribution. For example, if everyone in this sample had the same education, the variance in earnings would be reduced less than 4%-- even if prices remained unchanged. Thus equalization of earnings through compensatory education would require huge programs. Other compensatory programs may be more powerful, thought it is hard to specify such programs without having a good idea of what skills are rewarded in the market place. We approve of eliminating market imperfections but doubt that such a program would greatly reduce inequality, especially since the effects of most market imperfections are included in the common environment term. This leaves transfer programs.

FOOTNOTES

[1]Some or all of these issues are considered for example in Becker, Mincer, Champernowne, Lydall, Becker and Tomes, Meade, Taubman, Jencks and Sewell and Hauser.

[2]For one view of geneticists of why such estimates are not useful, see Feldman and Lewontin. For an alternative view, see Thoday.

[3]If some people are born or reared to be lazy, this criteria suggests that they are still entitled to transfer payments.

[4]We are ignoring mutations, which occur very rarely, i.e., about once in 100,000 or less. For discussion of the biological and statistical aspects of genes, see Cavalli-Sforza and Bodmer (1971).

[5]For a more complete discussion of the biological aspects, see Cavalli-Sforza and Bodmer (1971).

[6]Because these two equations incorporate supply and demand equations, the term "structural" is somewhat of a misnomer. The earnings equation, for example, under strong assumptions can be considered to be a hedonic price index in which prices for individual skills are the ones that equilibrate supply and demand curves for that skill. Alternatively, the coefficients of the right-hand side variables in this equation can be thought of as efficiency units weights for the inputs which are necessary for the production of skills. However, within the framework of our model we find it convenient to distinguish between the "structural" equations, which incorporate both observable and unobservable variables, and the "reduced form" equations, which are in terms only of the unobservable variables.

[7]The genetic indices are defined to be identical for each of the members of a MZ pair. The environmental indices are allowed to be less perfectly correlated across MZ or DZ pairs (although the correlation for the latter may differ from that for the former).

[8]The occupational status index utilized has its limitations in

representing these phenomenon. Nevertheless, it would seem to bear a definite relation at least to the first two of them. The degree of specific training and the extent of some nonpecuniary rewards (status, not having to punch a clock), for example, probably are positively correlated with the occupational index. Occupation-specific dummy variables, nevertheless, would permit a more compelling representation of such phenomenon.

[9]See, for example, Thaler and Rosen or Haspel and Taubman.

[10]However, the results for the partition of variance given later are essentially the same for schooling and occupational status or their lns.

[11]As shown in, say, Eaves, it is possible to express the genetic covariance between relatives as

$$r_n = c_1 c_2 \left[\frac{1+A}{2}\right]^n + 1/2^{(n+1)} A^{(n-1)} c_1 (1-c_2)$$

The parameters are defined as follows:

A is the correlation between the additive genetical deviations of spouses;

c_1 is the proportion of the total variation which can be ascribed to genetical differences (i.e., the "broad heretability");

c_2 is the proportion of the genetical variation which is additive;

n is the number of opportunities for genetical recombination in the shortest path in the pedigree linking the relatives under consideration (i.e., n=1 for sibs, 3 for first cousins).

With enough types of relatives it is possible to estimate these coefficients.

[12]There is some evidence supporting this view in Scarr Salapatek.

[13]Considering a twin pair as a unit of observation allows us to assume that disturbances are independently distributed.

[14]W' and W* are in turn defined in terms of W_o, W_D and W_M which correspond to Ω_o, Ω_D and Ω_M. However, we do not restrict W_o to be the same for MZ and DZ twins, but use instead the actual sample covariance, which we denote as W_o^M and W_o^D.

[15]This section is based on the material in Chamberlain, whom we also thank for correcting a major error in an earlier version.

[16]We are able to estimate σ_h^2 because Y_2 is not included in the structural equation for Y_4.

[17]The scale of unobserved variables is arbitrary.

[18]We eliminate pairs if either does not have earnings, whose derivation is described in Appendix B. About 100 pairs whose zygosity is unknown are not used in this analysis.

[19]The average 1973 earnings and education in our sample are $18,000 and 13 years. In the population as a whole, the corresponding figures for white veterans of the same cohort are about $15,500 and 12 years. About 1/4 of the differential can be eliminated if we reweight by parental education and region of birth so as to produce the average of white males born during the period 1917-1927 on these variables. See Appendix B.

[20]While our sample is not representative of the population, it seems likely that we have over and underrepresented various population groups rather than excluding all their members. For example, in Stauffer et al. (1950), it is indicated that there were huge differences in disqualification for mental problems by induction camp ranging from 1/2 of 1% to about 50%.

[21]However, since less than 5% of the sample have less than a ninth grade education, our results may not be appropriate for those with low education.

[22]In our statistical analysis, it is necessary to distinguish between MZ and DZ twins. For the most part, the twins' zygosity is determined by their answers to: "As children, were you and your twin alike as 'two peas in a pod' or of only ordinary resemblance?" This simple question assigns pairs accurately almost 95% of the time. See Appendix B, for details.

[23]See Taubman (1976), Appendix A, for discussion and references.

[24]For samples of this size, the 5% level of significance in an F test is about 1.2.

[25]See also the comparison below of our regression results with those based on Census data.

[26]A possible difficulty with the interpretation of the bias calculations occurs because of differential importance of measurement error in the various equations estimated. We consider this problem in Appendix A.

[27]See Mincer.

[28]The Chow test often is used for testing this type of null hypothesis. For the benefit of other users, we wish to report the following observation which we find disturbing. We calculated F statistics for alternative specifications of the same equation. In doing so we found that

the F ratio could rise from .8 to 1.6 in specifications that differed only in the inclusion of variables whose t statistics were less than 1.

[29]For the bias to vanish all the observed within pair variation must be measurement error.

[30]We can still estimate the G_2 coefficient in the OC_{67} equation because of the other restrictions in this model.

[31]Minimizing the negative of the log likelihood function is equivalent to maximizing the likelihood function itself. The minus sign on the function value is omitted in the tables.

[32]Alternatively, we know that σ_h^2 or $(1-\rho)$ is estimated from the Σ_M and Σ_D blocks. While the coefficients obtained from the Σ_M and Σ_D blocks will be affected by letting σ_h^2 differ from zero, the changes are not great.

[33]See Vandenberg.

[34]Goldberger has suggested that not allowing $\rho_{MZ} \neq \rho_{DZ}$ may be biasing our estimate of AM downwards. Such a bias reduces our estimate of genetic effects as would allowing ρ_{MZ} to be greater than ρ_{DZ} - though not necessarily by the same amount.

[35]See the material on identification in Section V above.

[36]We insert it here partly because we suspect that on-the-job-training and other environments would be much more alike for later occupations and earnings.

[37]In allocations of the total variance, we use the squares of these coefficients.

[38]When ρ is not restricted to 1, this coefficient drops to .005 but ρ is not significantly different from zero.

[39]One contributing factor may be that the status measure is generally related to average earnings of all persons in an occupation whatever their age. If those with less schooling enter high paying occupations later than those with more schooling, this result might occur.

[40]These are lower bound estimates since any differential treatment of twins in the family is included in the noncommon environment term.

APPENDIX A

John Bishop and several other economists have pointed out to us that the previous statements on bias must be qualified because of measurement error. As is well known if our true variable is s but we measure S which has a measurement error of v, then the bias from the measurement error (assuming v is uncorrelated with Y and with s) depends on σ^2_v / σ^2_s. Similarly in the within equations, the bias depends on σ^2_v / σ^2_s. We would expect $\sigma^2_{\Delta v} / \sigma^2_{\Delta s}$ to be greater than σ^2_v / σ^2_s since the brothers' true schooling will be correlated while the brother's measurement error either won't be correlated if we are dealing with wrongly reported numbers or less highly correlated if measurement error is expanded to include quality (denominated in units of earnings potential) as Welch has suggested.[1]

We can calculate, however, the effect of any measurement error variance on the estimates from the within and between estimates. The results (under the assumption that each estimate would be unbiased if there were no measurement error) which are given in Table 1 assume that the measurement errors across brothers are independent. As is evident in the table if independent measurement error for years of schooling is no greater than 10%, which seems large based on CPS-Census comparisons, the MZ within equations still imply a large bias.[2] Alternatively if the measurement error's variance is about 17 1/2% the within pair and individual estimates would both yield an estimate of about .093. If measurement error arises because of quality differences, which presumably are correlated across brothers, the MZ within biases will be smaller than those shown in Table A1.

Table A1

Estimates for Various Values of Measurement Error Variance

If	From MZ Within Equation	From Individual Equation	True Bias
$\sigma_v^2 = .05\sigma_s^2$	.032	.084	62%
$\sigma_v^2 = .10\sigma_s^2$	.048	.088	45%
$\sigma_v^2 = .15\sigma_s^2$	.070	.091	23%
$\sigma_v^2 = .20\sigma_s^2$	.121	.096	-26%

APPENDIX B

In 1955 a group of geneticists and medical researchers decided to assemble a "random" sample of white male twins to use in studying a wide variety of diseases. The sample is maintained by the National Academy of Science-National Research Council, who also control access by researchers. The sample construction and techniques are described in Jablon et al. (1967), from which the following quotation is taken:

> In 1955, experiments were initiated to explore methods of identifying twins who served in the Armed Forces during World War II. The method settled on was to obtain from the various state and city vital statistics offices in the U.S. copies of the birth records of all white male twins born in the years 1917-1927 and to match the names thus obtained against the VA Master Index (VAMI) to determine which twins survived with both entering military service. About 99% of all World War II veterans are represented in VAMI.
>
> It is not possible to tell just why the proportion of matches was so low. For a white male cohort born in 1920, about 86% survived to 1942. About 80% of the survivors served in the military forces in World War II, so that we might have expected to match 69% rather than 43.5%. Possible reasons for the discrepancy include higher mortality in the twins than in singletons born in the same year, higher rates of rejection for physical disability, and failures to match correctly at VAMI because of changes in name or inaccurate birth dates shown on the VAMI index card.[1]

Having received permission from the NAS-NRC to contact the twins, we mailed our survey on April 15, 1974 to 12,500 pairs for whom the NAS-NRC had recent addresses and who had cooperated with recent studies. On the first mailing, we received some 3650 valid replies and about 1000 returned because of wrong addresses.[2] In the second mailing of May 8, we included a special plea to those brothers whose twin had replied on the first mailing. We received 2400 responses from this mailing. We made a special mailing to those whose brother had replied and for whom we developed new addresses during June. Finally, on August 1 we made a registered mailing which included the same special plea and was restricted to those whose brother had previously responded.

In total, we received 6600 replies out of a possible 11,800 people with up-to-date addresses, even though on the third mailing we did not try to contact the nearly 6500 people where neither brother had previously responded. The 6600 replies contain 2468 matched pairs and 1600 unmatched individuals.

The restriction that both brothers served in the military may have some impact on the randomness of our sample with respect to the white male population born between 1917 and 1927. For example, this restriction excludes cases where one brother died before being eligible to serve in the military or was excluded from service because of mental and physical defects.[3] If his death or ineligibility were due to some genetic defect or poor family environment, it is likely that the survivor would have "poor" genes or environment and suffer from the same or related illnesses or defect, whether or not he was in the military. Since people were also likely to be rejected for service if they had already been convicted of a felony, we are likely to exclude too many criminals from our sample—given recidivism and the relatively high concordance rate on criminality for MZ and DZ twins. All these examples suggest that our sample had a truncated distribution of both genetic endowments and pre-adult environment. But it is not clear if the truncation is more severe with respect to G or N.

It is important that we correctly assign twin pairs as MZ or DZ. Except for a small portion of this sample for whom blood type and other pure genetic information is available, zygosity is assigned primarily on the basis of answers to the following two questions: "In childhood did your parents, brothers or sisters or teachers have trouble telling you apart?" and "As children were you and your twin alike 'as two peas in a pod' or only of ordinary family resemblance?" The pea question is the most important one for determining zygosity.[5] Cederlof et al. have made a detailed comparison of such questions with the assignments based on a number of purely genetic characteristics such as blood type, Rhesus factor, etc. For identical twins, the characteristics should be the same on all tests, and with enough tests, one can establish zygosity with as small a probability of error as desired. Among Swedish twins, Cederlof et al. found that the peas in a pod question agreed with the genetic information about 92% of the time with some of the error probably due to mistakes in the analysis of the chemical samples. Jablon et al. performed a similar analysis for a sub-sample of the NAS-NRC twins and concluded that the twins' self-assessment of zygosity was correct in 93% of the cases.

It can be shown that using data with such a misclassification will tend to make us overstate the noncommon environment variance and underestimate the genetic variance. A 5% misclassification error will cause the genetic effect to be understated by about 10% and a 10% misclassification will cause a 20% understatement.

Our sample is not a random drawing of the population in a number of respects. For example, Professor Hauser has made available to us the distributions by region of birth and parents' education of white male veterans born between 1917 and 1927. Our sample differs significantly from the population by having disproportionately higher parental education and disproportionately fewer people in the South.

As might be expected, there is also some evidence that there is a response bias in the sample. For example, some people who did not answer our questionnaire answered an earlier one which asked for occupation in 1967. The mean occupational status for people who answered this question but not our survey was about 80% of the mean for our responders. The cross sib correlations were also lower for the nonrespondents with the reduction greatest for MZ pairs.

Given the above discussion, it is not surprising to find that mean earnings and schooling in the sample are greater than in the population. In our sample earnings in 1973 are about $18,000. In 1969 for those 35 to 54 earnings for white males was about $12,000 (see the 1970 Census). Between 1969 and 1973 family income for this cohort increased by 21%. Assuming the same growth in earnings at all education levels, we estimate earnings in 1973 for the population to be about $15,000. In our sample, the respondents have about 1 1/2 years more education than in the population. We can recalculate what average 1973 population earnings would have been if the distribution of education in the population were the same as in our sample. The estimate is $16,500, which differs from our sample estimate by less than 10%. Since our regressions suggest that parental education has direct effects on earnings even given the twin's schooling and occupation, it seems likely that much of the remaining difference in average earnings could be eliminated by reweighing the sample.

We have completed some preliminary work in which we reweighting observations to reproduce the population percentages of father's education, mother's education and region of birth. The reweighted mean income eliminated about 1/4 of the difference between the sample mean and the original population mean (the latter was not adjusted directly for the twin's distribution of education). Interestingly this reweighted increases the variances and cross sib covariances for earnings proportionately.

Except for a rechecking in the original questionnaires and the replacement of one brother's no responses by the brother's response or sample mean if both brothers did not respond, there has been no editing of the education and occupation data. We have edited the 1973 earnings data as described below. The sample questionnaires will be found after that section.

The sample contains two different questions relating to the respondent's earnings. He was asked, "How much do you usually earn before deductions? Please indicate amount, then check pay period that applies to this amount: $_____ per hour _____, week _____, month _____, year _____, other (specify) _____." Then he was asked, "During 1973 how much did you earn?"

We constructed an annual earnings figure, denoted Y_A, from the first question by assuming a person worked 52 weeks or 12 months a year. The hourly pay rate was converted to a weekly figure by using the answer to the question, "How many hours do you normally work a week?" with no allowance for pay differentials for overtime.

The 1973 earnings, Y_{73} and Y_A need not be the same because (a) 1973 need not have been "usual" while Y_A is supposed to be based on "usual" earnings. (b) we did not collect data on number of weeks usually unemployed, (c) pay rates need not be constant between 1973 and the date the questionnaire was filled out in 1974--a period of rapid inflation, (d) there probably is recall error for 1973 earnings though most people filled out the questionnaire shortly after the usual deadline for filing income tax returns, (e) some people changed jobs or were promoted, and (f) some people only answered one of the 2 earnings questions. Despite all these obvious reasons for Y_{73} not to equal Y_A, it was felt that a comparison of the 2 series would aid both in establishing the accuracy of the figures and in editing the data. Since we would not expect the earnings figures to be exactly the same because of inflation and minor transitory events, we initially listed all cases in which Y_A fell outside of the range of plus or minus 15% of Y_{73}. Roughly 25% of the sample was listed but for about half this total either Y_A or Y_{73} was zero. Each of the remaining cases was examined by one of the authors (during a 2 day span for the preliminary tape and 4 days for the additional cases on the final tape) and in about 50 of these instances a change in Y_A or less frequently Y_{73} was made. In some cases it was clear that the wrong pay period was indicated on the tape. The most extreme example was a Y_{73} = \$22,000 and Y_A = (\$22,000 per week) (52 weeks) = \$1,440,000. For this and about 15 other observations, the pay period was altered. There were also a few instances in which it was apparent that Y_A or Y_{73} was off by a factor of 10 or 100. For cases that seemed very peculiar, we went back to the original questionnaires and found about 12 instances of incorrect coding. We also found several cases in which the person listed his known amount of pension with current earnings. We deleted pensions from earnings (but not income) whenever we learned of them.

Other changes were based on less information. There were several instances in which the nature of the job and answers to questions on extra work, overtime, and unemployment status indicate that the Y_{73} figure was more likely than the Y_A estimate. For example a fireman who often had a second job had a Y_{73} greater than Y_A, which was about \$10,000. The Y_A figure was raised to Y_{73}. In several other cases in which Y_A was substantially greater than Y_{73}, people had jobs that paid hourly and were in seasonal occupations (construction), or currently unemployed, or currently working far less than normal. For these people Y_A was set equal to Y_{73}. Conversely there were some hourly workers who normally worked say 48 hours and whose Y_{73} would approximately equal Y_A if time and a half for overtime were introduced. Once again Y_A was set equal to Y_{73}. Such judgmental changes as described in this paragraph were made to 1 1/2% of the observations.

Since some information is better than none and since people may not have felt the need to report earnings twice when the answers were the same, when we observed a positive estimate for Y_{73} and zero figure for Y_A,we substituted the Y_{73} estimate. Naturally, we also followed the reverse procedure. Including these changes, about 15% of the observations in Y_A and Y_{73} have been supplied or altered by the editing process.

Let us turn from a description of the editing process to the question of the accuracy of the data. If we assume that people did not bother to give Y_{73} because it was the same as the "usual" earnings, then for nearly 90% of our observations, the two estimates of earnings varied by no more than 15%. Given other information on the reliability of earnings data on mail surveys and on the transitory part of current earnings for these 90%, the earnings estimates are reasonable. Of course, there still are 10% of the sample whose two estimates differ by more than 15%. Many of these are self-employed. For these Y_A, which is generally reported on a monthly or annual basis, may be the more appropriate number to use though we received the impression that some of these people were giving their monthly draw which was less than normal business earnings. For another large group, it appeared that the difference depended upon the frequency of overtime, short work weeks and unemployment, but we could not determine which earnings estimate was the more representative. For this reason, we used both series in the preliminary analysis, which indicated that the ANOVA and regression results were fairly insensitive to the definition used.

APPENDIX C

NATIONAL RESEARCH COUNCIL
NATIONAL ACADEMY OF SCIENCES NATIONAL ACADEMY OF ENGINEERING

TWIN REGISTRY

FOR SCIENTIFIC PURPOSES...

For several years, research has been going on in connection with the effects of our environment on health. This questionnaire is part of an extensive study performed by the NATIONAL ACADEMY OF SCIENCES and supported by the U. S. PUBLIC HEALTH SERVICE. From the scientific point of view, it is of the greatest importance that every one to whom the questionnaire has been sent cooperate and answer it as carefully as possible.

Please note that your name does not appear on the questionnaire. The number on it is all the identification we need. The confidentiality of your replies will be fully respected.

First Here Are Some Questions Concerning Your General Health

1. Have you ever had any pain or discomfort in your chest?

- ☐ No → Proceed directly to Question 2
- ☐ Yes

a. When do you feel this pain or discomfort?

- ☐ When you are emotionally upset or excited
- ☐ When you walk fast or walk uphill
- ☐ When you walk at normal speed on level ground
- ☐ Under other circumstances

b. What do you do when you feel this pain or discomfort while you are walking?

- ☐ Stop walking or walk more slowly
- ☐ Take medicine and continue walking at the same speed
- ☐ Continue walking at the same speed without taking medicine

c. If you stop walking, regardless of whether you take medicine or not, how is the pain or discomfort then?

- ☐ The pain usually passes within ten minutes
- ☐ The pain usually continues for more than ten minutes

d. Where are the pains or the discomfort located?

- ☐ In the middle of the chest
- ☐ In the left side of the chest
- ☐ In the left arm
- ☐ In some other place

2. Do you get short of breath walking with other people at an ordinary pace <u>on the level</u>?

- ☐ No
- ☐ Yes → Do you get short of breath walking at your own pace?
  - ☐ No
  - ☐ Yes

3. Do you regularly or for extended periods of time have a cough?

- ☐ No → Proceed directly to Question 4
- ☐ Yes

a. For how many months in a row do you cough per year?

- ☐ Less than three months in a row
- ☐ More than three months in a row

b. For how many months in a row do you bring up phlegm from your chest?

- ☐ Less than three months in a row
- ☐ More than three months in a row

c. When is the cough worse?

- ☐ Wintertime
- ☐ During the other seasons
- ☐ It is equally troublesome throughout the year

4. Have you ever had a severe pain across the front of your chest lasting for a half hour or more?

- ☐ No
- ☐ Yes

5. Have you ever had a heart attack (coronary)?

- ☐ No
- ☐ Yes → What year?________
  If hospitalized, where?________

6. As a child, did you have croup?
- ☐ No
- ☐ Yes

7. Have you ever had asthma?
- ☐ No
- ☐ Yes → When did you have your last attack?
  - ☐ Within the past year
  - ☐ More than a year ago
  - ☐ Only as a child

8. Have you ever had hay fever, rose fever or allergic rhinitis (characterized by running nose, watery and itching eyes when you do not have a cold)?
- ☐ No
- ☐ Yes

9. Do you suffer from migraine or severe headaches?
- ☐ No
- ☐ Yes → Do you get symptoms before the headache starts which tell you that you will get a headache?
  - ☐ No
  - ☐ Yes → Do you then take medicine before the headache starts?
    - ☐ No
    - ☐ Yes

10. Did you have eczema when you were a baby?
- ☐ No
- ☐ Yes

11. Did you at times later in life have eczema-like skin conditions?
- ☐ No → Proceed directly to Question 13
- ☐ Yes

12. Do you know the name of the skin condition you have or have had?
- ☐ No
- ☐ Yes, psoriasis
- ☐ Yes, hives (urticaria)
- ☐ Yes, acne
- ☐ Yes, allergic rash
- ☐ Yes, eczema in knee or elbow fold
- ☐ Yes, allergic eczema
- ☐ Yes, others, namely ________

13. Do you have or have you ever had:
- ☐ Rheumatic fever or rheumatic heart disease
- ☐ Saint vitus dance (chorea)
- ☐ None of the above

14. When you become emotionally upset or are under emotional stress, do you often experience persisting (lasting more than an hour) disturbances in the form of:
- ☐ Pounding headache
- ☐ Palpitation of the heart
- ☐ Intestinal upset
- ☐ Sweating of palms
- ☐ None of the above

Next We Shall Ask You About Your Food Habits . . .

FOOD HABITS

15. How many times a day do you normally eat hot meals:
- ☐ At no time
- ☐ Once
- ☐ Twice
- ☐ Three times or more

16. Do you have to diet to keep your weight down?
- ☐ Yes
- ☐ No

17. How often do you eat pastries (coffee cake, sweet rolls, pie and cake)?
- ☐ Several times a day
- ☐ Once a day
- ☐ Less often

18. How often do you eat candy, candied fruit, chocolates?
- ☐ Several times a day
- ☐ Once a day
- ☐ Less often

19. Estimate your daily consumption of the following:
- ____ pieces of bread or rolls
- ____ glasses of milk or buttermilk
- ____ glasses of skim milk
- ____ cups of coffee, with ____ teaspoonfuls of sugar
- ____ cups of tea, with ____ teaspoonfuls of sugar

20. How often do you usually eat the following?

	Daily or almost daily	Once or twice a week	Once or twice a month	Less often
Pork (chops, ham, bacon, sausage, etc.)	☐	☐	☐	☐
Hot dogs, ground meats (meat loaf, hamburgers, etc.)	☐	☐	☐	☐
Roasts (other than pork), steaks, lamb, poultry	☐	☐	☐	☐
Dishes made of flour, cereals (dumplings, pancakes, spaghetti, macaroni, etc.)	☐	☐	☐	☐
Eggs	☐	☐	☐	☐
Fish and other seafood	☐	☐	☐	☐
Potatoes	☐	☐	☐	☐
Fruits and vegetables	☐	☐	☐	☐

21. Have you at any time made an extreme change in your food habits?
- ☐ No
- ☐ Yes → When? ________
  - In what way? ________
  - Why? ________

22. Are there any foods to which you are allergic?
- ☐ No
- ☐ Yes → What foods? ________
  - How do you react? ________

. . . About Your Smoking Habits . . .

SMOKING HABITS

23. Do you now smoke?
- ☐ No → Proceed directly to Question 30
- ☐ Yes

24. How many cigarettes do you usually smoke a day? ________ cigarettes

25. How many cigars do you usually smoke a day? ________ cigars

26. How many pipefuls of tobacco do you usually smoke a day? ________ pipefuls

27. If you now smoke cigarettes:

a. About how much do you inhale when smoking cigarettes?
- ☐ Do not inhale
- ☐ Inhale slightly
- ☐ Inhale moderately
- ☐ Inhale deeply

b. What type do you usually smoke?
- ☐ Filter-tip
- ☐ Without filter-tip

c. What brand do you usually smoke? ________

d. How old were you when you started smoking cigarettes? ________

28. If you now smoke cigars, about how much do you inhale when smoking cigars?
- ☐ Do not inhale
- ☐ Inhale slightly
- ☐ Inhale moderately
- ☐ Inhale deeply

29. If you now smoke a pipe, about how much do you inhale when smoking a pipe?
- ☐ Do not inhale
- ☐ Inhale slightly
- ☐ Inhale moderately
- ☐ Inhale deeply

30. If you <u>do not</u> smoke cigarettes now, did you ever smoke cigarettes regularly?
☐ No ☐ Yes

a. How long has it been since you last smoked cigarettes regularly? ________

b. How many cigarettes did you usually smoke per day? ________ cigarettes

c. How old were you when you started smoking cigarettes? ________

31. If you <u>do not</u> smoke cigars now, did you ever smoke cigars regularly?
- ☐ No
- ☐ Yes

32. If you <u>do not</u> smoke a pipe now, did you ever smoke a pipe regularly?
- ☐ No
- ☐ Yes

33. Do you chew tobacco or use snuff?
- ☐ Never
- ☐ Occasionally
- ☐ Regularly

. . . And About Your Consumption of Alcoholic Beverages . . .

DRINKING HABITS

34. **Have you at any time during the past year consumed beer, wine, or other alcoholic drinks (liquor)?**
- ☐ No →
- ☐ Yes ↓ Proceed directly to Question 35

Have you consumed any alcoholic beverage earlier in your life?
- ☐ No → Proceed directly to Question 46
- ☐ Yes

When did you stop drinking alcoholic beverages?

Year ________

Why did you stop?
- ☐ Fear of addiction
- ☐ Other reasons

____

And now proceed directly to Question 45 . . .

35. How often do you usually drink beer?
- ☐ Almost daily
- ☐ Once or twice a week
- ☐ Once or twice a month
- ☐ Once or twice a year
- ☐ Less often
- ☐ Never → Proceed directly to Question 37

36. On a day when you drink beer, how much do you usually drink?
- ☐ Less than one bottle or can (12 oz.)
- ☐ One bottle
- ☐ Two bottles
- ☐ Three bottles or more

37. How often do you usually drink wine?
- ☐ Almost daily
- ☐ Once or twice a week
- ☐ Once or twice a month
- ☐ Once or twice a year
- ☐ Less often
- ☐ Never → Proceed directly to Question 39

38. On a day when you drink wine, how much do you usually drink?
- ☐ A wine-glass or two
- ☐ A half bottle
- ☐ A half bottle – one bottle
- ☐ One bottle or more

39. How often do you usually drink liquor?
- ☐ Almost daily
- ☐ Once or twice a week
- ☐ Once or twice a month
- ☐ Once or twice a year
- ☐ Less often
- ☐ Never → Proceed directly to Question 41

40. On a day when you drink liquor, how much do you usually drink?
- ☐ Less than 1 shot (2 oz.)
- ☐ 1-2 shots
- ☐ 3-7 shots
- ☐ More than one pint (8 shots)

41. From time to time, one may have occasion to drink more than usual. How often do you drink alcoholic beverages in an amount that corresponds to at least one pint of liquor or two bottles of wine or four quarts of beer at one occasion?
- ☐ Daily or almost daily
- ☐ Once or twice a week
- ☐ Once or twice a month
- ☐ Less often
- ☐ Never

42. If you compare yourself with other persons you know, do you drink more or less alcoholic beverages than they do?
- ☐ Much more
- ☐ Somewhat more
- ☐ About the same as others
- ☐ Less than others

43. How often do you become really intoxicated?
- ☐ Daily
- ☐ Weekly
- ☐ Monthly
- ☐ Less often
- ☐ Never

44. How often do you have a "hangover?"
- ☐ Daily
- ☐ Weekly
- ☐ Monthly
- ☐ Less often
- ☐ Never

(Daily, Weekly, Monthly →) Do you then take another drink in the morning (an "eye-opener")?
- ☐ No
- ☐ Yes

45. Is there any alcoholic beverage to which you are allergic?
- ☐ No
- ☐ Yes → Which? ____________

How do you react? ____________

46. What is your height? ________ feet ________ inches

47. What is your weight? ____________ pounds

48. What was your weight at age 25? ____________ pounds

49. Have you checked your weight this year?
- ☐ Yes
- ☐ No

And Now Some Rather Specific Questions About Where You Have Lived Since the Second World War

50. For consecutive periods, fill in length of period, city or community, as well as state. Check also at the right of Table in what type of area you were living and working, respectively.

PERIOD OF TIME	CITY OR TOWN	STATE	LIVING IN			WORKING IN		
			Downtown Area	Suburban Area	Rural Area	Downtown Area	Suburban Area	Rural Area
1945 –			☐	☐	☐	☐	☐	☐
–			☐	☐	☐	☐	☐	☐
–			☐	☐	☐	☐	☐	☐
–			☐	☐	☐	☐	☐	☐
–			☐	☐	☐	☐	☐	☐
–			☐	☐	☐	☐	☐	☐
–			☐	☐	☐	☐	☐	☐
–			☐	☐	☐	☐	☐	☐

Questions Concerning Yourself and Your Family

51. Of the first 15 years of your life for how many years were you and your twin brother raised together?

No. of years ____________

52. During the greater part of your childhood (up to 15 years of age), did you live with your parents, or others?

- ☐ Yes, with father and mother
- ☐ With father only
- ☐ With mother only
- ☐ With other relative(s)
- ☐ With others

53. State precisely the nature of your father's main occupation throughout your childhood and adolescence (e.g., welder in shipyard, sales person in a bakery, manager of a department store):

____

54. State precisely the nature of your mother's main occupation throughout your childhood and adolescence:

____

55. Are you:

- ☐ Single ⟶ Proceed directly to Question 59.
- ☐ Married
- ☐ Remarried
- ☐ Divorced or separated
- ☐ Widowed

56. State precisely the nature of the main occupation of your wife:

____

57. Has she been employed during the greater part of your marriage?

- ☐ Yes
- ☐ No

58. Have you ever had children?

- ☐ Yes ⟶ How many? ____________
- ☐ No

59. Where did you live during the greater part of your childhood and adolescence?

- ☐ In a metropolitan area (population more than 1 million)
- ☐ In a large city (population between 100 thousand to 1 million)
- ☐ In smaller cities or other densely populated areas (population less than 100 thousand)
- ☐ In rural districts

60. How much physical exercise (outside of your work) have you had after 35 years of age?

- ☐ Hardly any
- ☐ Light exercise, e.g., regular walks, light gardening
- ☐ Minor sports (swimming, tennis, etc.)
- ☐ Hard physical training

61. Have you been employed or self-employed most of the time since age 25?

- ☐ No ⟶ Proceed directly to Question 69
- ☐ Yes

62. State precisely the nature of your main occupation (as in Question 53)

____

63. How many times have you changed employer since 25 years of age?

- ☐ 0 times
- ☐ 1 to 5 times
- ☐ 6 to 10 times
- ☐ More than 10 times

64. How often have you changed occupation after 25 years of age?

- ☐ 0 times
- ☐ 1 to 3 times
- ☐ More than 3 times

65. Has your work been mainly:

- ☐ Sedentary (involving mostly sitting)
- ☐ Moderately active
- ☐ Physically strenuous

66. In your work, have you been employed in a:

- ☐ Subordinate position
- ☐ Supervisory position (as a foreman, etc.)
- ☐ Self-employed

67. Do you often work overtime?

- ☐ No
- ☐ Yes, from time to time
- ☐ Yes, regularly

68. Have you often worked in addition to your regular employment?

- ☐ No
- ☐ Yes, from time to time
- ☐ Yes, regularly

69. What type of leisure activities, by and large, have you had in your adult life?

	frequently	some times	seldom or never
Home and family	☐	☐	☐
Visits with friends and relatives	☐	☐	☐
Home hobbies	☐	☐	☐
Movies, theater, art or music	☐	☐	☐
Reading newspapers, magazines, books	☐	☐	☐
Watching television, listening to radio	☐	☐	☐
Studies	☐	☐	☐
Club activities	☐	☐	☐
Extra work (overtime or outside employment)	☐	☐	☐
Home improvement and gardening	☐	☐	☐
Outdoor activities	☐	☐	☐
Participation in sports	☐	☐	☐

70. **Below are a number of statements regarding your occupational activities. Mark "X" the choice that corresponds to your opinion.**

STATEMENT	agree completely with statement	agree by and large with statement	am as much in favor as against statement	do not completely disagree	disagree completely
I have reached the position (within my vocation) to which I have aspired	☐	☐	☐	☐	☐
My ability and training have been used in full	☐	☐	☐	☐	☐
My training has not been adequate for the type of work I am doing	☐	☐	☐	☐	☐
My position involves too much responsibility	☐	☐	☐	☐	☐
I often have difficulty in finding enough time to complete the work assigned to me	☐	☐	☐	☐	☐
I have often felt somewhat uneasy in my work	☐	☐	☐	☐	☐
Mostly, I have gotten along well with my co-workers	☐	☐	☐	☐	☐
Financially, I have not achieved what I have hoped for	☐	☐	☐	☐	☐

Thank you for your help.

If our mailing address for you was incorrect, please indicate below your correct mailing address, including zip code.

Street ______________________________

City ________________ **State** ____________ **Zip** ________

We also want to bring up-to-date our address for your brother. If it is convenient, please indicate his address below.

Street ______________________________

City ________________ **State** ____________ **Zip** ________

Read carefully through the whole questionnaire again and make sure that nothing has been omitted.

FOOTNOTES

APPENDIX A

[1] $\sigma^2_{\Delta s} = 2\sigma^2_s - 2\sigma_{ss'}$ where prime indicates a sib's brother.

[2]The Census-CPS match may overstate measurement error since in the Census in some cases wives provide data for their husbands. This source of error is not found in our study.

APPENDIX B

[1]We have been told that inaccuracies in the VAMI index are no longer considered a major reason for the low match rate. Infant mortality was much higher for twins than for single births in the relevant time period. See Woodworth (1941).

[2]The mailing and processing of the questionnaire and the preparation of the data tapes was done by the Medical Follow Up Agency of the NAS-NRC, who performed these tasks most efficiently.

[3]For an indication of the reasons for and incidence of rejection, which was about 15% in World War II, see Stouffer et al., The American Soldier, Vol. IV. Incidentally, there is some evidence in this book that criteria for rejection differed widely by inductee camp and that the distributions are under-represented in the left-hand tail but not completely omitted.

[4]See Christiansen (1973).

[5]There is an adjustment factor based on the tell-apart question and imprecise genetic information such as ridge count on fingerprints.

REFERENCES

Becker, G., Human Capital, Columbia University Press, 1974.

Becker, G. and N. Tomes, "Child Endowments and the Quantity and Quality of Children," Journal of Political Economy, August 1976, vol. 84, pp. S143-S162.

Cavalli-Sforza, L. and W. Bodmer, The Genetics of Human Populations, W. H. Freeman & Co., 1971.

Chamberlain, G., "An Instrumental Variable Interpretation of Identification in Variable Components & MIMIC Models," 1977, this volume.

Chamberlain, G. and Z. Griliches, "Unobservables with a Variance Components Structure: Ability, Schooling and the Economic Success of Brothers," International Economic Review, 1975, vol. 16, pp. 422-449.

Champernowne, D. "A Model of Income Distribution," Economic Journal, 1953, vol. 63, pp. 318-351.

Christiansen, K. "Mobility and Crime Among Twins," International Journal of Criminology and Penology, 1973, Vol. 1, pp. 31-45.

Eaves, L., "Testing Models for Variation in Intelligence," Heredity, 1975, vol. 30, pp. 199-210.

Feldman, M. and Lewontin, R., "The Heritability Hang-Up," Science, 19 December, 1975, Vol. 193 pp.

Goldberger, A., "Structural Equation Models: An Overview," in Structural Equation Models in the Social Sciences, edited by A. Goldberger and O. Duncan, Seminar Press, 1973.

Haspel, A. and P. Taubman, "The Choice of Occupations: Some Empirical Evidence," SUNY-Buffalo, 1975.

Jablon, S. et al., "The NAS-NRC Twin Panel: Methods of Construction of the Panel, Zygosity Diagnosis and Proposed Use," American Journal of Human Genetics, 1967, vol. 19, pp. 133-161.

Jencks, C. et al., Inequality, Basic Books, 1972.

Meade, J., "The Inheritance of Inequality: Some Biological, Demographic, Social and Economic Factors," Proceedings of the British Academy, 1973, vol. 59, pp. 1-29.

Mincer, J., Schooling, Experience and Earnings, Columbia University Press, 1975.

Lydall, H., The Structure of Earnings, Oxford University Press, 1968.

Scarr Salapetek, S., "Twin Method: Defense of a Critical Assumption," University of Minnesota mimeo, 1975.

Sewell, W. and R. Hauser, Education, Occupation and Earnings: Achievement in the Early Career, Academic Press, 1975.

Stauffer, S. et al., The American Soldier, Vol. 4, Princeton University Press, 1975.

Taubman, P., "The Determinants of Earnings: Genetics, Family, and Other Environments, A Study of White Male Twins," American Economic Review, December 1976.

_______________, Sources of Inequality of Earnings, North-Holland, 1975.

Thaler, R. and S. Rosen, "Estimating the Value of a Life," in Household Production and Consumption, N. Terleckyj ed., Columbia University, Press, 1976.

Thoday, J., "Genetics and Educability," Journal of Biological Education, 1972, vol. 6, pp. 323-329.

Vandenberg, S. "Assortive Mating or Who Marries Whom," Behavior Genetics, Vol. 2, pp. 127-157, June 1972.

Woodworth, R. Heredity and Environment, Social Science Research Council Bulletin, 47, New York, 1941.

Zellner, A., "Estimation of a Regression Relationship Containing Unobservables," International Economic Review, vol. 11, pp. 441-454,1970.

CHAPTER 4

MORE ON BROTHERS*

Gary Chamberlain and Zvi Griliches

I. Introduction

In an earlier paper (Chamberlain-Griliches, 1975) we developed a model for and attempted to estimate the effects of an unobserved ability variable on the observed income-schooling relationship across individuals. The major features of that paper were the assumption that the unobservable had a variance-components structure and the use of data on brothers (from Gorseline, 1932) to estimate its effects. For that data set the usual estimate of returns to schooling was changed very little when we allowed for the presence of such an unobservable variable.[1]

The findings of the earlier paper were subject, however, to several reservations. First, they related to a rather old and non-representative set of data. Second, our model relied on the presence of a second (indicator) equation, one for occupational success, to identify the parameters of interest in the income-schooling relationship. There is some question, however, whether it is legitimate to treat income and occupation as two different measures of success. Third, since no test scores had been collected in that data set, our interpretation of the unobservable as "ability" was rather tenuous. Moreover, there appeared to be very little commonality in the occupation experience of brothers at that time and place (Indiana in the late 1920's), the intra-class correlation being only about .04. This is rather unfortunate for a model that relies heavily on the additional occupational equation and the difference between family and individual effects for identification.[2] All this lead us to pursue such questions on other, potentially better, data sets.

In this paper we will report the first results of our analysis of data on pairs of brothers from the National Longitudinal Survey of Young Men.[3] This is a national sample of (originally) over 5,000 young men, interviewed first in 1966 and followed up annually through 1971, and biannually thereafter.[4]

To save on sampling costs the Census Bureau based this sample and the parallel samples of the Older Men, Mature Women, and Young Women

---

*We are indebted to the NIE and NSF for financial support and to Bronwyn Hall and Stephen Messner for research assistance.

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 in the Young Men panel. We will be using the 1966-69 data for a subset of pairs of these brothers in the analysis below. We intend, in the near future, to extend our study to brother and sister pairs and to family groups of three and larger.

The NLS data set for young men (and young women) is of interest to us because, besides the usual economic and demographic variables, it contains also a test of the "knowledge of the world of work," which was administered at the time of the initial interview (1966), and IQ type test scores which were collected from the high schools of the respondents.[5] We shall interpret the first test (KWW) as a test of "late" ability and the second (IQ) as a test of "early" ability. The availability of such test scores allows us to dispense with the rather dubious occupational success equation of our earlier model and to relate the unobservable variable(s) closer to what is commonly meant by "ability." Moreover, the availability of two test scores and of additional information on the family structure leads to a significantly over-identified model, allowing us to test some of its more dubious restrictions and to explore the estimation of significantly more general models.

No data set is perfect, however. From our point of view, the basic difficulty with this sample, besides many missing entries, is the extreme youth of the respondents. As of 1969 close to half of the total sample was still in school. 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. Moreover, brothers, because they had to be living at home in 1966 to be so defined, are even younger. (See Table 1 for details). Hence, it is hard to interpret their current status as a good indication of their ultimate success in life. But in addition to their current status, respondents were also asked about their expected ultimate educational attainment and their expected ("desired at age 30") occupation.[6] We have scaled (valued) their (three-digit) occupational expectations by the median earnings of all U.S. males in 1959 in these occupations, converting them thereby into an "expected" income concept.

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

1. It allows us to deal with expected income and schooling 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.
2. It comes close to dealing with the ex-ante optimizing behavior of individuals, as discussed in the Becker (1967) or Rosen (1973) models, uncontaminated by the ex-post encounter with reality.
3. Most important, it allows us to triple our sample of brothers, avoiding 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:

1. We are dealing with expectations and not "reality."

2. 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.
3. The causality from schooling to earnings is much less clear for expectational variables.[7]

Nevertheless, we believe that the expectational data are of intrinsic interest and that the advantages enumerated above outweigh the disadvantages. In any case, we intend to check our results against the "real" data for the "not-enrolled in 1970" sub-sample in the near future.[8]

II. Models, Variables, and Hypotheses

The major focus of this study is on estimating the income-schooling relationship in the presence of an unobserved ability variable. Since we have test scores for most of the individuals and also information on their family relationship, the question arises how to utilize all this information most effectively. In doing so we have to recognize that the available test scores are only fallible proxies for the desired "ability" concept, that the later test score (KWW) is itself affected by schooling, and that the available background information on the families of these brothers does not really describe their "environments" fully or adequately.

The model outlined below (in Figure 1) tries to set up a general framework which would allow the consideration of most of the issues raised by such data. It consists of two separate equations for the early (IQ) and later (KWW) test scores, allowing the second one to depend also on the level of schooling attained at the time of the survey (1966); two schooling equations, one for schooling as of 1966 (S66) and one for the expected ultimate schooling attainment (ES);[9] and an equation for the logarithm of expected income (EY). All of these endogenous variables are related to a general set of exogenous variables $\underset{\sim}{X}$ (such as age, dummy variables for the dating of the expectational variables, and regional variables), measured family background variables $\underset{\sim}{B}$ (such as race, father's occupation, mother's education, number of siblings, and a "culture" index), and an unobserved ability variable (A), having a variance-components structure across family members.

Figure 1 outlines the details of the model. Several points should be emphasized here. (1) We are assuming, provisionally, that there is only one common left out variable A. A two-common factor generalization of this model will be considered below. (2) We are assuming that measured (and unmeasured) family background variables all work via the unobserved family component (f) of A, except for the schooling equations, where we allow family "wealth" to have a separate effect from family "ability." In particular, we assume that these variables do not enter the income equation directly. That is a testable hypothesis (in some versions of our model). (3) The only source of cross-equation correlations is the left out ability variable A. All the other disturbance components (the u's) are mutually uncorrelated. This too is testable in some versions of our model.

Table 1: Characteristics of the NLS Brothers Sample
N = 584 (292 pairs)

Variable	Means and Standard Deviations Total	Within	ρ
Age 69	20.3		
	2.3	1.4	.23
ES	14.8		
	2.3	1.1	.51
S66	11.3		
	1.7	1.1	.15
EY	8.67		
	.404	.270	.11
KWW	34.9		
	7.7	4.5	.32
IQ	102.8		
	15.9	7.5	.56
FOMY 14	5418		
	2179		
BLACK	.20		
CULTURE	2.5		
	.76		
SIBLINGS	3.6		
	2.1		
SMSA	.67		
ROS	.32		

The lower number in a pair of numbers is the standard deviation.

ρ - One minus twice the ratio of the within families variance to the total variance.
ES - Expected total schooling to be completed eventually, in years.
S66 - Schooling completed in 1966, in years.
EY - Logarithm of the 1959 median earnings (in dollars) of all males in the occupation expected (desired) at age 30.
KWW - Score on the "knowledge of the world of work" test, administered in 1966.
IQ - Score on IQ type tests, collected from the high school last attended by the respondent.
FOMY14 - Occupation of father or head of household when respondent was 14, scaled by the median earnings of all U.S. males in this occupation in 1959.
CULTURE - Index based on the availability of newspapers, magazines and library cards in the respondent's home.
SMSA - Respondent in SMSA in 1969.
ROS - Respondent in South when 14.

Table 2: Correlations Between Selected Variables

584 Individuals

Variables	ES	EY	KWW	IQ
ES	1.000	...	...	...
EY	.415	1.000	...	...
KWW	.324	.220	1.000	...
IQ	.482	.275	.482	1.000

Cross-Sib Correlations

292 pairs

Variables	ES'	EY'	KWW'	IQ'
ES	.508	...	...	...
EY	.228	.109	...	...
KWW	.267	.196	.323	...
IQ	.333	.204	.359	.555

Figure 1

NLS Brothers: General Model

a. Structural Model

$$
\begin{aligned}
y_1 &= IQ = \underset{\sim}{X}\underset{\sim}{\alpha}_1 && +\lambda_1 A + {}_1 \\
y_2 &= S66 = \underset{\sim}{X}\underset{\sim}{\alpha}_2 + \underset{\sim}{B}\underset{\sim}{\delta}_2 && +\lambda_2 A + u_2 \\
y_3 &= KWW = \underset{\sim}{X}\underset{\sim}{\alpha}_3 + \beta_3 y_2 && +\lambda_3 A + u_3 \\
y_4 &= ES = \underset{\sim}{X}\underset{\sim}{\alpha}_4 + \underset{\sim}{B}\underset{\sim}{\delta}_4 + \beta_4 y_2 && +\lambda_4 A + u_4 \\
y_5 &= EY = \underset{\sim}{X}\underset{\sim}{\alpha}_5 + \beta_5 y_4 && +\lambda_5 A + u_5 \\
A_{ij} &= \underset{\sim}{B}\underset{\sim}{\delta}_1 + f_i + g_{ij}
\end{aligned}
$$

b. Reduced form residuals e (net of X and B):

$$
\begin{aligned}
e_1 &= \lambda_1(f+g) && +u_1 \\
e_2 &= \lambda_2(f+g) && +u_2 \\
e_3 &= (\lambda_3+\beta_3\lambda_2)(f+g) && +u_3 + \beta_3 u_2 \\
e_4 &= (\lambda_4+\beta_4\lambda_2)(f+g) && +u_4 + \beta_4 u_2 \\
e_5 &= \{(\lambda_5+\beta_5(\lambda_4+\beta_4\lambda_2)\}(f+g) && +u_5 + \beta_5(u_4+\beta_4 u_2)
\end{aligned}
$$

c. Variance-covariance Matrix of the Reduced Form Residuals:

$$
\underset{n\to\infty}{\text{plim}}\ \underset{\sim}{R} = E(\underset{\sim}{e}_{ij}\ \underset{\sim}{e}'_{ij}) = \underset{\sim}{\Theta} + \underset{\sim}{\Sigma}
$$

Family components of variance : $\underset{\sim}{\Theta} = \underset{\sim}{d}\underset{\sim}{d}'$, with

$$\underset{\sim}{d} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 + \beta_3\lambda_2 \\ \lambda_4 + \beta_4\lambda_2 \\ \lambda_5 + \beta_5(\lambda_4 + \beta_4\lambda_2) \end{bmatrix} \sigma_f$$

Individual components of variance: $\underset{\sim}{\Sigma} = \tau\underset{\sim}{\Theta} + \underset{\sim}{V}$

Individual "ability" related components of variance: $\tau\underset{\sim}{\Theta}$, with $\tau = \sigma_g^2/\sigma_f^2$

Other individual components of variance:

$$\underset{\sim}{V} = \begin{matrix} \sigma_1^2 & 0 & 0 & 0 & 0 \\ & \sigma_2^2 & \beta_3\sigma_2^2 & \beta_4\sigma_2^2 & \beta_4\beta_5\sigma_2^2 \\ & & \sigma_3^2+\beta_3^2\sigma_2^2 & \beta_3\beta_4\sigma_2^2 & \beta_3\beta_4\beta_5\sigma_2^2 \\ & & & \sigma_4^2+\beta_4^2\sigma_2^2 & \beta_5(\sigma_4^2+\beta_4^2\sigma_2^2) \\ & & & & \sigma_5^2+\beta_5^2(\sigma_4^2+\beta_4^2\sigma_2^2) \end{matrix} ,$$

assuming that $Eu_k\ u_h = Egu_k = 0$ for all $k \neq h$

d. Variance-covariance matrix of average family residuals:

$$\bar{e}_{ik} = \frac{1}{p}\,\Sigma_j e_{ijk}$$

$$\underset{n\to\infty}{\text{plim}}\ \underset{\sim}{\bar{R}} = E(\underset{\sim}{\bar{e}}_i\underset{\sim}{\bar{e}}_i') = \underset{\sim}{\Theta} + \frac{1}{p}\underset{\sim}{\Sigma}$$

Table 3

NLS Brothers: Estimates of a One-Factor Model

Dependent Variable	Other Variables In Equation	Coefficients of S66	ES	A	Estimated σ_u
Y_1 = IQ	$\underset{\sim}{X}, \underset{\sim}{B}$			9.03 (.16)	9.10
Y_2 = S66	$\underset{\sim}{X}, \underset{\sim}{B}$			.423 (.063)	.90
Y_3 = KWW	$\underset{\sim}{X}$	.651 (.295)		2.87 (.31)	5.64
Y_4 = ES	$\underset{\sim}{X}, \underset{\sim}{B}$	.368 (.096)		.895 (.135)	1.59
Y_5 = EY	$\underset{\sim}{X}$		.057 (.009)	.045 (.016)	.363

$A_{ij} = \underset{\sim}{B}_i\delta_1 + f_i + g_{ij}, \sigma_f^2 + \sigma_g^2 \equiv 1, \quad \tau = \sigma_g^2/\sigma_f^2 = \underset{(.217)}{.723}, \quad \sigma_g^2/\sigma_A^2 = .420.$

Estimated asymptotic standard errors in parenthesis

$\underset{\sim}{X}$: Age, ROS, Dates (of the expected variables)

$\underset{\sim}{B}$: FOMY14, MED, Number of Siblings, Black, Culture

TABLE 4

Variance Decompositions of the Major Dependent Variables

	Due to the variance of						
	$\underset{\sim}{B}$	f	g	u_1	u_2+u_4	u_3	u_5
IQ	.32	.20	.14	.34			
KWW	.18	.11	.08		.01	.62	
ES	.26	.13	.09		.52		
EY	.06	.04	.03		.05		.82

The contribution of $\underset{\sim}{B}$ is "total", both direct and indirect via the tests and schooling.

Table 5

Estimated Effects of Background Variables

	FOMY 14 in \$1000	MED in years	Black	Culture Index: 0-3	Sibs Number
IQ: $\lambda_1\underset{\sim}{\delta}_1$ (in IQ points)	.98	.71	-11.2	5.1	-.38
ES (in years) via G: $(\lambda_4+\beta_4\lambda_2)\underset{\sim}{\delta}_1$	.11	.08	- 1.30	.59	-.04
direct: $\underset{\sim}{\delta}_4+\beta_4\underset{\sim}{\delta}_2$	.07	.13	2.16	.19	.01
ES Total	.18	.21	.86	.78	-.03

The effects via S66 are solved out and included in ES

Also (4), this version of our model does not allow for the possible endogeneity of schooling itself or for possible errors in its measurement. The two-factor model to be discussed below will go part way towards relaxing these assumptions.

The first part of Figure 1 (Panel A) lists the major equations of the model and defines the variance-components structure of the ability variable, allowing its family component to be correlated with the measured family background variables ($\underset{\sim}{B}$). The second panel (b) of Figure 1 exhibits the residuals of the reduced form equations of this system, which we get by solving out the endogenous schooling variables y_2 and y_4 and by sweeping out the exogenous and at this point unconstrained $\underset{\sim}{X}$ and $\underset{\sim}{B}$ variables. The third panel (c) defines the variance-covariance matrix of these reduced form disturbances, labels and exhibits its components, dividing it into three pieces: the contribution of the family related component of the ability variable $\underset{\sim}{\theta}$, the contribution of the individual components of the ability variable $\tilde{\tau}\underset{\sim}{\theta}$ and the contribution of the rest of the individual components of variance-$\underset{\sim}{V}$. The last panel (d) defines the variance-covariance matrix of averaged (over family members) reduced form residuals and exhibits its relationship to the previously defined components.

It is clear that as defined the model is heavily overidentified. Let $\underset{\sim}{R}$ be the sample covariance matrix of the residuals from the reduced form equations estimated by ordinary least squares, and let $\bar{\underset{\sim}{R}}$ be the sample covariance matrix of the averaged residuals, averaged separately over each family. Given $\underset{\sim}{R}$ and $\bar{\underset{\sim}{R}}$, we can get an estimate of $\underset{\sim}{\theta}$ from $\hat{\underset{\sim}{\theta}} = \frac{p}{p-1}[\bar{R} - \frac{1}{p} R]$ and an associated estimate of $\hat{\underset{\sim}{\Sigma}}$ from $\hat{\underset{\sim}{\Sigma}} = \underset{\sim}{R} \, \hat{\underset{\sim}{\theta}}$.[10]

Since by hypothesis $\underset{\sim}{V}$ contains several zeros, we have several estimates of τ to choose from (since $\underset{\sim}{\Sigma} = \tau \underset{\sim}{\Theta} + \underset{\sim}{V}$ and we already have an estimate of $\underset{\sim}{\Theta}$). Given a $\tilde{\tau}$ we can get $\tilde{V}$ which in turn gives us three estimates of β_5: v_{25}/v_{24}, v_{45}/v_{44} and v_{35}/v_{34}. A maximum likelihood procedure for combining all these estimates (and other restrictions) is described in our earlier paper. Given two test scores, our model is significantly more overidentified than was the model we used to analyze the Gorseline brothers data. Some of these extra restrictions will be traded away below to allow us to relax several of the more questionable assumptions of this model. However, before we complicate it further, we shall present and discuss the results of fitting the simpler version to our data.

III. Results: One-Factor Model

The results of estimating the model outlined in Figure 1 are summarized in Table 3.[11] We do find a significant common left out factor which connects all these variables, but its direct contribution to the income equation is rather small (λ_5^2 Var (A) is about .004 as compared to the variance of EY of .163 or the estimated $\sigma^2_{u_5}$= .13). The total estimated contribution, via both schooling and tests, is of course larger. It is given in Table 4, which presents the variance decompositions implied by these results. Table 5 lists the coefficients of the background variables in the ability and schooling equations. The estimated schooling coefficient $\hat{\beta}_5$ = .057 is similar to what Griliches (1976) found for the total sample of NLS Young Men (N = 3025). There, the TSLS estimates of β_5 were between .052 and .057 depending on whether IQ or KWW were the test variables used in the equation. We can also compare it to the least squares estimates that we would get if we assumed that the IQ variable is not subject to error. In our sample of brothers, $b_{EY, ES\,.\,IQ}$ is .066 overall and .067 in the "within" portion of the sample (based only on differences between brothers, i.e., on $\underset{\sim}{\Sigma}$ only).

These estimates are based on (at least) two substantive restrictions: the exclusion of the background variables from having a direct effect on expected income (except through the ability variable) and the restriction to only one common left-out variable. The first restriction appears to be supported by the data. Allowing the variables to enter the EY equation in an unconstrained fashion reduces $\hat{\sigma}_{u_5}$ from .363 only to .362 and raises the estimated β_5 to .059. Testing the restriction that the $\underset{\sim}{B}$ variables have the same relative effects on EY (net of ES) as on IQ gives an insignificant $\chi^2(5)$ = 5.5, indicating that our data do indeed come close to satisfying this restriction. On the other hand, testing the one-factor model against a two-factor model pushes us to extend our model and estimates to the much more complex two-factor case. The test is based on the canonical correlations between the reduced form residuals and a set of family indicator dummy variables. As shown in Chamberlain and Griliches (1975), the squared canonical correlations have a null value of .5 (for p = 2) under the hypothesis of no family structure. With only $\underset{\sim}{X}$ in the reduced form, the squared canonical correlations are (.80, .69, .57, .57, .48). Partialling on $\underset{\sim}{X}$ and $\underset{\sim}{B}$ gives (.69, .62, .56, .56, .48). Testing the second largest squared canonical correlation vs. .5 gives $\chi^2(4)$ = 44.2 (net of X), and partialling on $\underset{\sim}{X}$ and $\underset{\sim}{B}$ we still get a very significant $\chi^2(4)$ = 16.2.

IV. The Two-Factor Model

By adding another unobservable factor to our model, we can allow for a much more complex structure of data generation. In particular, we can allow for the possibility of (a) two different types of "human capital," (b) components of ability that affect tests and schooling but not income, and (c) errors in our measures of schooling. Unfortunately, we cannot really distinguish between them.

Let us rewrite the previous "ability" measure A as being made up of two variables, G and H; i.e., substitute $\lambda G + \gamma H$ for λA in Figure 1. Each of these variables is assumed to depend on the observable family background variables $\underset{\sim}{B}$ and to have a family components variance structure for its unobservable part with $G_{ij} = \underset{\sim}{B}_i \underset{\sim}{\pi} + f_i + g_{ij}$ and $H_{ij} = \underset{\sim}{B}_i \underset{\sim}{\eta} + m_i + h_{ij}$. f and m are the parts of the family components of G and H that are uncorrelated with the observed background variables $\underset{\sim}{B}$. The family (f and m) and individual (g and h) unobservable components may be correlated across the two factors with correlation coefficients r_{fm} and r_{gh} respectively. We have thus added eight new parameters to our original model (5 γ's, 2 r's, and $\tau_2 = \sigma^2_h / \sigma^2_m$). That this results in a severe identification problem should come as no surprise to us.

The identification of this model is explored in greater detail in the Appendix.[12] We show there that the β's remain identified in this model but that the λ's and γ's cannot be estimated separately without imposing additional substantive assumptions. That is, we can estimate the schooling coefficient in the presence of two unobservable factors, but we cannot really interpret the separate contributions of such factors uniquely.

Table 6 presents the results of the two-factor version of our model based on the normalizations $\gamma_1 = 0$ and $r_{gh} = 0$. This assumes that H is not IQ and that only the common family components (f and m) may be correlated with each other. Alternative interpretations will be discussed shortly.

Allowing for a second unobservable factor raises the estimated schooling coefficient from .057 to .064 without changing much else in the model. This is similar to what happens in the larger individuals sample when schooling is also allowed to be endogenous though the change in the coefficient is smaller here (see Griliches, 1976a for details). It is consistent, as we shall show below, with several interpretations of the role of ability and schooling and the possibility of errors in the measures of both variables.

Given the maximum likelihood estimates in Table 6, we can reinterpret the factor structure by imposing alternative normalizations. The β's and σ's are unaffected by the choice of normalization. One possibility (given as alternative 1 in Table 7) is to set $\gamma_1 = \lambda_5 = 0$ and let r_{gh} be free. This implies (in terms of the estimates of Table 6) that H* = H + (.020/ .0079)G. H* is now interpreted as the relevant initial human capital variable and G as the "true score" on the IQ test. Both variables have positive effects on KWW, but H* has a negative partial effect on ES, as does H in Table 6. This corresponds to interpreting H*, the unobserved initial human capital, as consisting of two pieces: a part (G) that reduces the cost of schooling in terms of the time and effort required to complete a grade, and a part (H) that does not affect the cost of schooling, except via foregone earnings. Then for a given level of G, increases in H (or H*) are transmitted to the schooling equation with a negative sign. This is because the optimum total stock of human capital is unaffected by H, so that an increase in initial H implies that less additional investment in schooling is needed to attain the fixed target.

For an endogenous schooling model the basic optimality condition is $\frac{dY}{dS} / i = Y + C(G)$; i.e., the discounted increment to life time earnings should equal the foregone income Y plus the monetary equivalent C of the

Table 6

NLS Brothers: Estimates of the Two-Factor Model

Dependent Variable	Coefficients (and standard errors) of S66	ES	G	H	Estimated σ_u	Other variables in equation
IQ			10.60 (.97)		7.41	$\underset{\sim}{X}$, $\underset{\sim}{B}$
S66			.374 (.163)	-.376 (.75)	.844	$\underset{\sim}{X}$, $\underset{\sim}{B}$
KWW	1.01 (.31)		2.38 (.74)	.232 (.07)	5.71	$\underset{\sim}{X}$
ES	-.33 (.501)		1.07 (.30)	-1.79 (.20)	.05	$\underset{\sim}{X}$, $\underset{\sim}{B}$
EY		.064 (.009)	.020 (.021)	.0079 (.0039)	.362	$\underset{\sim}{X}$

$\sigma_f^2 + \sigma_g^2 = \sigma_m^2 + \sigma_h^2 \equiv 1$ $\quad \tau_1 = \sigma_g^2/\sigma_f^2 = 1.05$ (.46) $\quad \tau_2 = \sigma_h^2/\sigma_m^2 = 2.39$ (1.75)

$r_{fm} = .005$ $\quad r_{gh} \equiv 0$

Table 7

Alternative Two-Factor Structures

Dependent Variable	Alternative 1 Coefficients of G	H*	Alternative 2 Coefficients of G*	$\tilde{H}$
IQ	10.60		9.37	3.92
S66	1.32	-1.02	.15	.49
KWW	1.79	.63	2.19	.65
ES	5.61	-4.87	.09	2.07
EY		.02	.02	

$\sigma_f^2+\sigma_g^2 = \sigma_{m*}^2+ \sigma_{h*}^2 \equiv 1$ $\quad \sigma_{f*}^2+ \sigma_{g*}^2 = \sigma_{\tilde{m}}^2+ \sigma_{\tilde{h}}^2 \equiv 1$

$\tau_1 = 1.05$ $\quad \tau_2^* = 1.16$ $\quad \tau_1^* = 1.16$ $\quad \tilde{\tau}_2 = 1.95$

$r_{fm*} = .95$ $\quad r_{gh*} = .91$ $\quad r_{f*\tilde{m}} = .32$ $\quad r_{g*\tilde{h}} \equiv 0$

differential utility cost of foregone leisure required to complete a grade. Given the assumed semi-log functional form, the optimality condition is $\beta_5/i = 1 + C(G)/Y$. Since C depends only on G, an increase in H (or H*) for given G does not affect the optimum Y, and so it must be offset by a fall in S. Note the implication that the reduced form for Y should not depend on S. This last implication is not borne out very well in our results, since the estimate of $\beta_5(\gamma_4 + \beta_4\gamma_2)$ is -.107, which more than offsets the estimated $\gamma_5 = .008$.

Since in Table 5 $r_{fm} \simeq 0$ and $r_{gh} = 0$, we have $r_{fm^*} = .95$ and $r_{gh^*} = .91$, implying that G and H* are much the same variables; hence we have quite a bit of difficulty in distinguishing their separate effects clearly.

An alternative, and possibly more substantive reinterpretation is based on reversing the sign of H and setting $G^* = G-(.0079/.020)H$ so that $\gamma_5 = 0$. We also set $\tilde{H} = H + \psi G^*$, where ψ is chosen so that $r_{g^*\tilde{h}} = 0$. Now (Alternative 2 in Table 7) both factors have non-negative signs in all the equations. The interpretation of the increase in $\hat{\beta}_5$ over the one-factor estimate is that the true score consists of both G* and $\tilde{H}$, but only G* is relevant for earning income. So when we use IQ as a proxy for G*, $\tilde{H}$ is part of the "error" and enters the income residual with a negative sign. But $\tilde{H}$ is positively correlated with ES, leading to a downward bias in $b_{EY,ES.IQ}$.

Several additional extensions and simplifications are possible. In Table 6 we see that τ_2 is not significantly different from zero. When $\underset{\sim}{B}$ is unconstrained, the ML estimate of τ_2 is in fact zero. Setting τ_2 to zero and leaving $\underset{\sim}{B}$ unconstrained gives $\hat{\beta}_5 = .063$. Leaving out the second factor also from KWW does not affect the fit significantly ($\chi^2(1) = 1.56$) but raises $\hat{\underset{\sim}{\beta}}_5$ further to .067. Adding a third family factor leads to no change or improvement in the rest of the model. The test statistic for the presence of a third family factor is based on testing the third largest squared canonical correlation against .5. With unconstrained background variables the test gives $\underset{\sim}{\chi}^2(3) = 4.1$. Thus, a simplified version of our model would contain one factor (G) made up of both family and individual components and a second purely family factor (m) with a correlation of about $r_{fm} = .35$ between the two unobservable family components. Since this version is not really any easier to interpret than the one given originally in Table 6, we will not pursue it further here. Nor, given the arbitrariness in defining G and H, will we present variance decompositions for the various alternative versions of these models. We do want to conclude, however, that (a) adding a second unobservable factor raises the schooling coefficient, and (b) there is no evidence that a third family factor would be required to rationalize these data.

Discussion

Since we have both within and between family data (on expected schooling, income, and IQ) and two test scores, we can allow for a variety of possible biases in our data. Consider first the possibility that the left-out variables are pure family effects, such as family wealth or connections (without any other errors or dependencies in the data). Then the appropriate estimate of the schooling coefficient is the one derived from $\underset{\sim}{\Sigma}$ (the "within" variance-covariance matrix): $b^{W}_{EY,ES\,.\,IQ} = .067$. If the only problems are errors of measurement in the test scores (and possibly also in schooling), then the appropriate estimate of β_5 can be had from $\hat{\theta}$[the smoothed one- or two-factor (rank) version of $\underset{\sim}{B}$ (between) - $1/2\underset{\sim}{W}$ (within)]. Computing it from an unsmoothed $\underset{\sim}{\theta}$ without taking out the background variables (i.e., assuming that they work only via schooling and IQ), we get $b^{\theta}_{EY,ES.\,IQ} = .061$. Allowing for both, unspecified left out family factors and errors in test scores, would lead us to estimate β_5 from the within families data ($\underset{\sim}{\Sigma}$) using one test score as an instrument for the other. For our data, this yields a $\hat{\beta}_5$ of .055. But this estimate ignores all the between family variation and allows for no errors in the schooling variable or dependencies between the disturbances in the schooling and income equations. In contrast, our one- and two-factor models described earlier allow for the effects of unobservables at both the family and individual levels. The one-factor model does not allow for the endogeneity of schooling and gives essentially the same results as the instrumental variables-within-$\underset{\sim}{\Sigma}$ estimator (.057 vs .055). The two-factor model relaxes this restriction and leads us back to an estimate of β_5 of around .064. Since we start with an overall least squares estimator of $b_{EY,\,ES.IQ,\underset{\sim}{X}} = .066$ (holding age, dates, and ROS constant), we never really wander too far away from it. To put it differently, since much of our methods depends on utilizing and capitalizing on the different information contained in the within and between families portions of our sample, starting out with a sample where the two rather different data cuts give almost the same answers (.066 and .067 respectively) provides little leverage for our methodology.[13]

This may explain also why we get rather different answers than those reported by Cochrane-Jencks-Olneck for the Kalamazoo and Talent brothers set and by Behrman-Taubman for the NRC twins data. All of these studies report rather sharp differences between their total and within estimators of the schooling coefficients. Since we get little difference to start out with, we cannot extrapolate the observed difference much further.

This is not the place to discuss the other studies in any detail. We'd like to note only that the Cochrane-Jencks-Olneck estimates do not allow for differences in age or work experience among their brothers. It was shown by Mincer (1974) and Griliches (1976, 1977) that the omission of the experience variable can affect greatly the estimated schooling coefficients (especially at younger ages).

Postscript on Expectations

Our analysis has not been tailored specifically to data on expectations. The principal problem is that expected earnings and expected schooling are likely to be jointly determined. This section models the joint determination of these variables by treating them symmetrically.

Assume a causal relationship connecting the actual variables:

(1) $Y = \beta S + \lambda A + u.$

So earnings are determined by schooling, ability, and a random disturbance (u) that is uncorrelated with S and A. The disturbance can be interpreted as luck or unanticipated random events which occur after the completion of schooling. But this interpretation of the disturbance is much more tenuous with data on expectations. Future random events that are unanticipated cannot possibly influence expected earnings (EY) or expected schooling (ES); events that are anticipated are likely to influence both EY and ES. Such events are like characteristics of the individual, and their omission can bias the results. Although schooling may be a determinant of earnings, expected schooling does not cause expected earnings. Rather they are jointly determined in that an individual may choose an income level and then calculate how much schooling is needed to attain it.

A simple formulation of the problem can be based on the assumption of conditionally unbiased expectations. The assumption is that

(2) $E(Y \mid EY, ES, A) = EY$

$E(S \mid EY, ES, A) = ES$

This means that the expectation errors (Y-EY and S-ES) are uncorrelated with the expectations and with ability. So the expectation errors have a mean of zero even within a group of individuals who all have the same expectations and ability. We also assume that

(3) $E(u \mid EY, ES, A) = 0,$

since the interpretation of u in the realization model is that it is unpredictable from information that is available when the expectations are formed. Combining these conditional expectations with (1) gives:

(4) $EY = \beta ES + \lambda A.$

So this formulation results in a symmetric treatment of EY and ES by dispensing with the "luck" disturbance altogether.

Since the restriction that $\sigma_u = 0$ is quite strong, it may be preferable to allow for random departures from the assumption of unbiased expectations. A possible specification is

(5) $EY = \tilde{Y} + \epsilon, ES = \tilde{S} + \delta,$

where $\tilde{Y}$ and $\tilde{S}$ satisfy the consistency condition in (4). Then provided that the errors ϵ and δ are not themselves correlated across brothers, the main implication of the model is to allow the individual specific disturbances in the ES and EY equations to be correlated with each other.[14] So in Table 1 we would have $E(u_4 u_5) \neq 0$.

Consider the following model:

$$
\begin{aligned}
(6)\quad IQ_{ij} &= && \lambda_1 A_{ij} && && + u_{1ij} \\
S66_{ij} &= && \lambda_2 A_{ij} && + \gamma_2 m_i && + u_{2ij} \\
KWW_{ij} &= \beta_3\, S66_{ij} && + \lambda_3 A_{ij} && && + u_{3ij} \\
ES_{ij} &= \beta_4\, S66_{ij} && + \lambda_4 A_{ij} && + \gamma_4 m_i && + u_{4ij} \\
EY_{ij} &= \beta_5 ES_{ij} && + \lambda_5 A_{ij} && && + u_{5ij}.
\end{aligned}
$$

This model augments the one-factor model of Table 1 by introducing a purely family specific factor (m_i) which only enters the schooling equations. It is interpreted as "opportunity" [Becker (1967)], reflecting the extent to which the family subsidizes the cost of schooling or provides encouragement for schooling achievement.

This model is identified (except for β_4) even in the extreme case in which all of the individual disturbances (u_1, ..., u_5) are freely correlated. Then $\underset{\sim}{\Sigma}$ contains no information about the structural slope parameters. The identification of these parameters must be based completely on $\underset{\sim}{\theta}$. The key to studying identification in $\underset{\sim}{\theta}$ is that the $\underset{\sim}{\theta}$ moments are determined solely by the family specific effects (with i subscripts). This gives

$$
\begin{aligned}
(7)\quad IQ_i &= && + \lambda_1 f_i \\
S66_i &= && \lambda_2 f_i + \gamma_2 m_i \\
KWW_i &= \beta_3 S66_i && + \lambda_3 f_i \\
ES_i &= && \bar{\lambda}_4 f_i + \bar{\gamma}_4 m_i \\
EY_i &= \beta_5 ES_i && + \lambda_5 f_i ,
\end{aligned}
$$

where $\bar{\lambda}_4 = \lambda_4 + \beta_4 \lambda_2$ and $\bar{\gamma}_4 = \gamma_4 + \beta_4 \gamma_2$.

Since we are assuming that the error in IQ as a measure of A is not correlated across brothers, we can treat IQ as a perfect measure of f (up to a scale factor). So β_5 is identified by a regression in $\underset{\sim}{\theta}$ of EY on ES and f. A key identifying restriction is that m_i does not enter directly the structural relationship connecting EY, ES and f. This restriction follows from the exclusion of m from the realization equation in (1). The argument for the exclusion is that opportunities (m) do not affect earnings given ability (f) and investment (S) in human capital.

The $EY = \beta_5 ES + \lambda_5 f$ relationship captures the joint determination of EY and ES since they are both correlated with f, and there is no additional purely family residual that is correlated with EY but not with ES. EY and ES are treated symmetrically in $\underset{\sim}{\Sigma}$ also, since there the EY and ES residuals are freely correlated.

We have fit this model by maximum likelihood with the $\underset{\sim}{X}$ variables of Table 1 included freely in all of the equations. We have not included the observed background variables ($\underset{\sim}{B}_i$) since their exclusion from the EY equation is one of the identifying restrictions. We could have included $\underset{\sim}{B}$ as determinants of f and m, but this just gives a decomposition of the variance of f and m into a part that is explained by $\underset{\sim}{B}$ and a residual variance, without altering the estimates of the other parameters.

The estimation procedure amounts to first fitting a rank 2 approximation to the unconstrained estimate $\hat{\underset{\sim}{\theta}} = \frac{p}{p-1}(\bar{\underset{\sim}{R}} - \frac{1}{p}\underset{\sim}{R})$.

The rank 2 approximation can be obtained from the solution to a canonical correlation problem as shown in Chamberlain and Griliches (1975). Then we use the constrained estimate of the $\underset{\sim}{\theta}$ moments to compute a regression of EY on ES and IQ. Because of the rank restriction, we can obtain precisely the same estimate of β_5 by taking the reciprocal of the EY coefficient in the regression of ES on EY and IQ.[15] Thus the joint determination of EY and ES is reflected in our symmetric treatment of these variables.

With the u's freely correlated, the estimate of β_5 is .056 with a standard error of .02.[16] We also fit the model allowing for correlation only between u_2 and u_5 and between u_4 and u_5. This resulted in $\hat{\beta}_5 = .054$ with a standard error of .02. The estimated correlation between u_4 and u_5 is only .12, and the decline in goodness of fit relative to the free $\underset{\sim}{\Sigma}$ model is not significant, with a likelihood ratio test giving $\chi^2(6) = 7.6$. In fact, constraining all of the u's to be uncorrelated also gives an insignificant decline in fit relative to the free $\underset{\sim}{\Sigma}$ model: $\chi^2(8) = 7.8$. In this version the point estimate of β_5 stays about the same, but there is a considerable increase in the precision of the estimate: $\hat{\beta}_5 = .061$ with a standard error of .008.

We have also allowed a third purely family factor to play the role of u, entering only the EY equation and uncorrelated with everything but EY. The estimated variance of this third factor is essentially zero and completely insignificant, which is consistent with our finding that the canonical correlations indicate that a rank 2 approximation to $\underset{\sim}{\theta}$ is

Table 8: NLS Not-Enrolled (1970), Working Brothers: Estimates of a Two-Factor Model

Dependent Variable	Coefficients (and standard errors) of S66	SC	XBT	A	m	Estimated σ
IQ				10.96 (1.40)		7.73
S66				.655 (.350)	.436 (.139)	.949
KWW	.618 (.527)			2.62 (1.00)		6.09
SC	.839 (.164)			.406 (.279)	.365 (.145)	.839
XBT		.046 (.003)		-.0051 (.0059)		.068
LW		.072 (.020)	-.365 (.278)	-.0049 (.028)		.335

$\sigma_f^2 + \sigma_g^2 \equiv 1$ $\quad \sigma_m^2 \equiv 1$ $\quad \tau_1 = \sigma_g^2/\sigma_f^2 = .993$ (.523) $\quad \tau_2 = \sigma_h^2/\sigma_m^2 \equiv 0$

$r_{fm} = -.38$ (.51)

161 pairs; age, region, dates ($\underset{\sim}{X}$), and measured family background ($\underset{\sim}{B}$) variables enter unconstrained in all of the equations.

SC - years of schooling completed in 1970 (in years).

XBT - $e^{-.1 \cdot EXP70}$, where EXP70 is cumulated work experience in 1970 (in years). See Griliches (1976a) for more details.

LW - logarithm of hourly earnings on the current or last job in 1970.

adequate.[17] We conclude that a more careful treatment of the joint determination of expected earnings and expected schooling does not alter the central tendency of our results, although the precision of the estimates declines.

Postscript on the Earnings of the Not-Enrolled Subsample

We originally chose to use the expectation data because the extreme youth of the sample creates fewer problems with expected occupation at age 30 than with the current wage rate. Moreover, the inclusion of those still enrolled in school tripled our sample of brothers. This section compares the expectation results with the estimates based on the recently available data for 161 pairs of brothers who were not enrolled in 1970. This data also allows us to use the actual hourly wage rate instead of the median earnings in the individual's occupation.

The principal problem due to the extreme youth of the sample is that current earnings are a poor indicator of future prospects. Much of the labor force behavior of these youths is characterized by search, experimentation, and often a lack of "seriousness." Hence we expect schooling and ability to have weaker effects on current than on subsequent earnings. Even for the group that has settled into its long-run trajectory, on-the-job training drives a wedge between potential and actual earnings. If schooling and ability are complementary to this training (and not substitutes for it), then omitting an adequate measure of training will bias the schooling and ability coefficients to zero.

Table 8 presents our estimates from the 1970 data for a constrained two-factor model. Now the second factor is family specific ($\tau_2 = 0$) and only enters the schooling equations. It is interpreted as family wealth or family encouragement for schooling achievement. Allowing this opportunity factor to have a direct effect on earnings gave a small and insignificant effect. But an attempt to allow for within family variation in the second factor did not converge; the residual schooling variance (σ_4^2) was tending to zero, whereas a positive variance is necessary for the identification of the model (see Appendix).

The main results are the negligible ability coefficient and the substantial schooling coefficient in the earnings equation.[18] We expect the ability coefficient to increase as the sample ages, as in Sewell and Hauser's (1975) study of Wisconsin high school seniors. The finding of an already substantial schooling coefficient is borne out in Griliches's study based on over 2000 youths in the NLS survey.

On the whole we feel that the results from the expectation data are better indicators of the eventual peak schooling and ability effects. There will be an opportunity to check this as the panel is resurveyed every two years. More important, we will eventually be able to estimate models that combine the expectation data with the realizations.

FOOTNOTES

*We are indebted to the NIE and NSF for financial support and to Bronwyn Hall and Stephen Messner for research assistance.

[1]See Griliches (1974, 1977) for surveys of related literature and Chamberlain (1975) for more details on the model.

[2]This was pointed out to us by C. Jencks.

[3]See Chamberlain (1974 and 1976a) for analyses of related data sets not based on family structure.

[4]See Griliches(1976) and Parnes et al.(1964-74) 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 oversampled in a 3 to 1 ratio. The original sample consisted of 5,225 individuals of whom 3,734 were white. By 1969 about 23 percent of the original sample was lost, 13 percent of it only temporarily (to the Army).

[5]Unfortunately, the IQ test scores are unavailable for about a third of the sample, including all those who did not continue school beyond the 9th grade.

[6]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 is 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 that did not originate from the 1969 survey (DATELOMY and DATE 66, identified collectively as DATES).

[7]This point is developed in a postscript to the paper.

[8]Some results on the not-enrolled subsample are presented in the second postscript.

[9]Clearly the presence of S66 in the ES equation does not have a structural interpretation. It is simply a convenient substitute for a non-zero correlation between the S66 and ES residuals (u_2 and u_4). The S66 equation also lacks a structural interpretation. It is simply an auxiliary regression that summarizes the correlation between S66 and A. Moreover,

we can dispense with S66 by solving it out of the KWW and ES equations. Then the new ES and KWW residuals are correlated; but the correlation can be reset to zero by allowing ES to enter the KWW equation. This version of the model is still identified--simply substitute IQ as a proxy for G in the EY equation and use the other brother's IQ score as an instrumental variable—and it yields very similar estimates of the earnings equation.

[10] p is the number of brothers per family. In our case p=2. This estimate is not efficient since it doesn't take into account the implication of the model that $\underset{\sim}{\theta} = \underset{\sim}{d}\underset{\sim}{d}'$ is of rank 1.

[11] These and subsequent estimates were computed using an amended ACOVSM (1971) program. The standard errors were estimated by perturbing the solution and interpolating from the resulting likelihood ratio χ^2 (chi-square) statistics.

[12] The Appendix treats the model in which S66 has been solved out of the KWW and ES equations with the resulting correlation between their residuals captured by including ES in the KWW equation. Thus the S66 equation, which is difficult to interpret, is not necessary for our analysis, and in fact the estimates are very similar without S66. This model is also covered by the general identification analysis in Chamberlain (1976b).

If we do use S66, then there is a very simple argument for identification: a combination of IQ, KWW, and S66 will serve as a proxy for $\lambda G + \gamma H$ in the EY equation, with a measurement error that depends on μ_1 and μ_3; then the IQ and KWW scores for the other brother can be used as instrumental variables.

[13] Without including IQ in the equation, the difference is a bit larger: $b^T_{EY, ES.\underset{\sim}{X}} = .074$ versus $b^W_{EY, ES.\underset{\sim}{X}} = .069$, but still quite small.

[14] See Chamberlain (1976c) for some details on this model and for a discussion on combining the expectations with the eventual realizations.

[15] We can obtain the same estimate of β_5 by regressing EY on ES, KWW and S66. Once again the reciprocal regression also gives the same estimate.

[16] Note that τ is not identified in this version of the model. Also the correlation between f and m is not identified.

[17] We have already seen (in the concluding section of the body of the paper) that using an unconstrained $\underset{\sim}{\theta}$ results in $b^{\theta}_{EY,ES.IQ} = .061$.

[18] The OLS estimates give $b_{LW,IQ.SC,XBT,\underset{\sim}{X},\underset{\sim}{B}} = .0001$ with a standard error of .001 and $b_{LW,SC.IQ,XBT,\underset{\sim}{X},\underset{\sim}{B}} = .076$ with a standard error of .014.

APPENDIX

Identification of the Two-Factor Model

We will work with the following two-factor model:

$$T_1 = \quad + \lambda_1 G + \gamma_1 H + u_1$$

$$S = \quad \lambda_2 G + \gamma_2 H + u_2$$

$$T_2 = \beta_3 S + \lambda_3 G + \gamma_3 H + u_3$$

$$Y = \beta_4 S + \lambda_4 G + \gamma_4 H + u_4,$$

where $G_{ij} = f_i + g_{ij}$, $H_{ij} = m_i + h_{ij}$,

with f and m freely correlated and g and h freely correlated. The u's are assumed to be uncorrelated with G, H and with each other. It is straightforward to allow for observable background characteristics and to allow the late test (T_2) to depend upon only a part of S.

We will show that β_3 and β_4 are identified. The λ's and γ's, however, are not identified without additional normalizations. For example, we can set $\lambda_1 G + \gamma_1 H = \lambda_1 G^*$ and $\lambda_n G + \gamma_n H = \lambda_n G^* + (\gamma_n - \lambda_n \frac{\gamma_1}{\lambda_1}) H$ for n = 2, 3, 4. This transformation of the factor structure implies that $\gamma_1 = 0$. But the β's are not affected by this transformation; hence showing that they are identified with the $\gamma_1 = 0$ normalization will show that the β's are identified in the seemingly more general model. We shall see that an additional normalization, as well as two scale normalizations, is required in order to fix the factor structure. One possibility is $\lambda_4 = 0$.

The reduced form is:

$$\underset{\sim}{y} = \underset{\sim}{D} \begin{bmatrix} G \\ H \end{bmatrix} + \underset{\sim}{\varepsilon},$$

with $\underset{\sim}{y}' = (T_1, S, T_2, Y)$,

$$\underset{\sim}{D} = (\underset{\sim}{d}, \underset{\sim}{k}), \underset{\sim}{d} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 + \beta_3 \lambda_2 \\ \lambda_4 + \beta_4 \lambda_2 \end{bmatrix}, \underset{\sim}{k} = \begin{bmatrix} 0 \\ \gamma_2 \\ \gamma_3 + \beta_3 \gamma_2 \\ \gamma_4 + \beta_4 \gamma_2 \end{bmatrix}, \varepsilon = \begin{bmatrix} u_1 \\ u_2 \\ u_3 + \beta_3 u_2 \\ u_4 + \beta_4 u_2 \end{bmatrix}$$

We allow for unconstrained correlations between f and m and between g and h by projecting m onto f and h onto g:

$$m = \eta_1 f + m^*, \quad h = \eta_2 g + h^*,$$

where m* is uncorrelated with f and h* is uncorrelated with g. The scale normalizations are that $\sigma_f = \sigma_{m^*} = 1$. The implicit assumptions that $\sigma_f \neq 0$ and $\sigma_{m^*} \neq 0$ are necessary for identification. Then with $\underset{\sim}{C}_1 = \begin{bmatrix} 1 & 0 \\ \eta_1 & 1 \end{bmatrix}$, $\underset{\sim}{C}_2 = \begin{bmatrix} 1 & 0 \\ \eta_2 & 1 \end{bmatrix}$, we have $\underset{\sim}{y} = \underset{\sim}{D}\underset{\sim}{C}_1 \begin{bmatrix} f \\ m^* \end{bmatrix} + \underset{\sim}{D}\underset{\sim}{C}_2 \begin{bmatrix} g \\ h^* \end{bmatrix} + \underset{\sim}{\varepsilon}$.

Let $\underset{\sim}{\theta}$ be the covariance matrix generated by the family effects (f, m*) and let $\underset{\sim}{\Sigma}$ be the matrix of within family covariances. Then we have

$$\underset{\sim}{\theta} = \underset{\sim}{D}\, \underset{\sim}{C}_1\, \underset{\sim}{C}'_1\, \underset{\sim}{D}'$$

$$\Sigma = \underset{\sim}{D}\, \underset{\sim}{C}_2\, \Phi\, \underset{\sim}{C}'_2\, \underset{\sim}{D}' + V,$$

where $\underset{\sim}{\Phi} = \begin{bmatrix} \tau_1 & 0 \\ 0 & \tau_2 \end{bmatrix}$ with $\tau_1 = \sigma_g^2/\sigma_f^2$, $\tau_2 = \sigma_{h^*}^2/\sigma_{m^*}^2$, and

$$\underset{\sim}{V} = E(\underset{\sim}{\varepsilon}\,\underset{\sim}{\varepsilon}') = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ & \sigma_2^2 & \beta_3\sigma_2^2 & \beta_4\sigma_2^2 \\ & & \sigma_3^2 + \beta_3^2\sigma_2^2 & \beta_3\beta_4\sigma_2^2 \\ & & & \sigma_4^2 + \beta_4^2\sigma_2^2 \end{bmatrix}$$

The identification problem is to determine which of the structural parameters can be obtained from $\underset{\sim}{\theta}$ and $\underset{\sim}{\Sigma}$.

Let $\underset{\sim}{D}\,\underset{\sim}{C}_1 = (\underset{\sim}{b}, \underset{\sim}{k})$ where $\underset{\sim}{b} = \underset{\sim}{d} + \eta_1 \underset{\sim}{k}$, and note that $k_1 = 0$. Therefore the first column of $\underset{\sim}{\theta}$ is $b_1\underset{\sim}{b}$. This will identify $\underset{\sim}{b}$ up to a sign normalization provided that $b_1 \neq 0$, i.e., if $\lambda_1 \neq 0$. Then $\underset{\sim}{\theta} - \underset{\sim}{b}\,\underset{\sim}{b}' = \underset{\sim}{k}\,\underset{\sim}{k}'$ will identify $\underset{\sim}{k}$.

In $\underset{\sim}{\Sigma}$ we set $\underset{\sim}{D}\,\underset{\sim}{C}_2\,\underset{\sim}{\Phi}^{1/2} = (\underset{\sim}{s}, \sqrt{\tau_2}\,\underset{\sim}{k})$, where $\underset{\sim}{s} = \sqrt{\tau_1}\,(\underset{\sim}{d} + \eta_2\underset{\sim}{k})$. Then $k_1 = 0$ together with the zero off-diagonal elements in the first row of $\underset{\sim}{V}$ imply that the first row of $\underset{\sim}{\Sigma}$ identifies $\underset{\sim}{p}' = s_1\,(s_2, s_3, s_4)$. Since $\underset{\sim}{s}$ is a linear combination of $\underset{\sim}{d}$ and $\underset{\sim}{k}$, we have $\underset{\sim}{p} = \alpha_1 \bar{\underset{\sim}{b}} + \alpha_2 \bar{\underset{\sim}{k}}$, where $\bar{\underset{\sim}{b}}$, $\bar{\underset{\sim}{k}}$ contain the last three elements of $\underset{\sim}{b}$, and $\underset{\sim}{k}$. This uniquely determines α_1, α_2 provided that $(\bar{\underset{\sim}{b}}, \bar{\underset{\sim}{k}})$ has rank 2. This will be true as long as the reduced form coefficients of G and H in the last three equations are not proportional to each other. Then with $\underset{\sim}{c} = \alpha_1 \underset{\sim}{b} + \alpha_2 \underset{\sim}{k}$, we can identify $\underset{\sim}{s}$ up to a scale factor from $\underset{\sim}{s} = \underset{\sim}{c}/s_1$. We can solve for s_1 provided that T_1 is correlated within families with at least one other variable. For example, if $\sigma_{12} \neq 0$ then $\sigma_{12} = s_1 s_2 = c_1 c_2/s_1^2$ implies that $s_1^2 = c_1 c_2/\sigma_{12}$.

Then given $\underset{\sim}{s}$ we have $\underset{\sim}{\Sigma} - \underset{\sim}{s}\,\underset{\sim}{s}' = \tau_2\,\underset{\sim}{k}\,\underset{\sim}{k}' + \underset{\sim}{V}$.

We will write the lower right 3 x 3 block of this set of equations as

$$\underset{\sim}{\Psi} = \tau_2\,\underset{\sim}{t}\,\underset{\sim}{t}' + \begin{bmatrix} \sigma_2^2 & \beta_3\sigma_2^2 & \beta_4\sigma_2^2 \\ & \sigma_3^2 + \beta_3^2\sigma_2^2 & \beta_3\beta_4\sigma_2^2 \\ & & \sigma_4^2 + \beta_4^2\sigma_2^2 \end{bmatrix},$$

where $\underset{\sim}{t}' = (k_2, k_3, k_4)$ can be obtained from $\underset{\sim}{\theta}$. For a given value of τ_2, we can solve for

$$\sigma_2^2 = \psi_{11} - \tau_2 t_1^2$$

$$\beta_3 = (\psi_{12} - \tau_2 t_1 t_2)/\sigma_2^2$$

$$\beta_4 = (\psi_{13} - \tau_2 t_1 t_3)/\sigma_2^2 .$$

But we can also solve for

$$\beta_3 \beta_4 = (\psi_{23} - \tau_2 t_2 t_3)/\sigma_2^2 .$$

This yields:

$$\tau_2 = (\psi_{23}\psi_{11} - \psi_{12}\psi_{13}) / (\psi_{11} t_2 t_3 + \psi_{23} t_1^2 - \psi_{13} t_1 t_2 - \psi_{12} t_1 t_3).$$

So once we have obtained $\underset{\sim}{s}$, the model reduces to the S, T_2, Y equations with one common factor (H), which is equivalent to the model in Chamberlain and Griliches (1975).

The rank condition is that the denominator in our solution for τ_2 must not equal zero. Expressing this condition in terms of the structural parameters gives:

$$\sigma_m^2 * \sigma_2^2 \gamma_3 \gamma_4 \neq 0 .$$

So H must have a family component, and there must be variation in S that is independent of G and H. Also T_2 must be related directly to H (not just via its dependence on S). T_2 is not identified if $\gamma_4 = 0$. But then H is not contributing to the bias in the OLS estimate of β_4 ; hence it is easy to show that β_4 and λ_4 (as well as λ_1 and λ_2) are identified without using the T_2 equation at all.

As for the other parameters, it is clear that $\underset{\sim}{k}$ together with β_3 and β_4 will identify $\underset{\sim}{\gamma}$. Also $b_1 = \lambda_1$ and $s_1 = \sqrt{\tau_1}\ \lambda_1$ will identify λ_1 and τ_1. $\eta_1 - \eta_2$ is identifiable from $\underset{\sim}{b} - \underset{\sim}{s}/\sqrt{\tau_1} = (\eta_1 - \eta_2)\underset{\sim}{k}$. But λ_2, λ_3, λ_4 and η_1 (or η_2) are not identified without an additional normalization. The problem is that we cannot distinguish our model from one that sets $\lambda_4 G + \gamma_4 H = \gamma_4 H^*$ and $\lambda_i G + \gamma_i H = (\lambda_i - \frac{\gamma_i \lambda_4}{\gamma_4}) G + \gamma_i H^*$, i = 2, 3. An additional normalization could be $\lambda_4 = 0$. Then $b_4 = \eta_1 \gamma_4 + \beta_4 b_2$ and $\underset{\sim}{d} = \underset{\sim}{b} - \eta_1 \underset{\sim}{k}$ will identify η_1 and $\underset{\sim}{\lambda}$. Note that the ML estimates of β_1, β_2 and $\underset{\sim}{\gamma}$ are not affected by this additional normalization.

REFERENCES

Becker, G.S., 1967. "Human Capital and the Personal Distribution of Income," Woytinski Lecture No. 1. Ann Arbor: University of Michigan.

Berhrman, J.R. and P. Taubman, 1976. "Intergenerational Transmission of Income and Wealth," American Economic Review, Vol. 66(2), pp. 436-440.

Chamberlain, G., 1977. "Education, Income and Ability Revisited," Journal of Econometrics, Vol. 5(2), pp. 241-257. Reprinted in D. Aigner and A. Goldberger, Latent Variables in Socio-Economic Models, Amsterdam: North-Holland Publishing Company, 1977.

____________, 1975. "Unobservables in Econometric Models," unpublished Ph.D. dissertation, Harvard University.

____________, 1976a. "Unobservables in Earnings Functions: Applications in the Malmo Data," Harvard University, unpublished.

____________, 1976b. "Identification in Variance Components Models," Harvard Institute of Economic Research Discussion Paper No. 486.

____________, 1976c. "Expectations, Realizations, and Measurement Error," Harvard University, unpublished.

Chamberlain, G., and Z. Griliches, 1975. "Unobservables with a Variance Components Structure: Ability, Schooling and the Economic Success of Brothers," International Economic Review, Vol. 16(2), pp. 422-449. Reprinted in D. Aigner and A. Goldberger, eds. Latent Variables in Socio-Economic Models, Amsterdam: North Holland Publishing Company, 1977.

Corcoran, M., C. Jencks and M. Olneck, 1976. "The Effects of Family Background on Earnings," American Economic Review, Vol. 66(2), pp. 430-435.

Griliches, Z., 1974. "Errors in Variables and Other Unobservables," Econometrica, Vol. 42(6), pp. 971-998.

__________, 1976. "Wages of Very Young Men," Journal of Political Economy, Vol. 84(4), pp. S69-S85.

__________, 1977. "Estimating the Returns to Schooling: Some Econometric Problems," Econometrica, Vol. 45(1), pp. 1-22.

Gorseline, D.E., 1932. The Effect of Schooling Upon Income. Bloomington: Indiana University Press.

Joreskog, K.G., G.T. Gruvaeus and M. v. Thillo, 1971. "ACOVSM: A General Computer Program for Analysis of Covariance Structures," Research Bulletin RB-711, Princeton Educational Testing Service.

Mincer, J., 1974. Schooling, Experience, and Earnings. New York: NBER.

Parnes, H.S. et al, 1970-73. Career Thresholds, Vols. 1-4, Center for Human Resource Research, Ohio State University, Columbus, Ohio.

Rosen, S., 1973. "Income Generating Functions and Capital Accumulation," Harvard Institute of Economic Research Discussion Paper No. 306, Cambridge, MA.

Sewell, W.H. and R.M. Hauser, 1975. Education, Occupation and Earnings: Achievement in the Early Career. New York: Academic Press.

CHAPTER 5

ON THE USE OF SIBLING DATA TO ESTIMATE THE EFFECTS OF FAMILY BACKGROUND, COGNITIVE SKILLS, AND SCHOOLING: RESULTS FROM THE KALAMAZOO BROTHERS STUDY

Michael R. Olneck
Institute for Research on Poverty and
Department of Educational Policy Studies
University of Wisconsin-Madison

Introduction

During the last ten years, sociologists have devoted considerable effort to measuring and modelling the effects of family background on the economic attainments of men (Blau and Duncan, 1967; Duncan, Featherman, and Duncan, 1972; Jencks et al., 1972; Sewell and Hauser, 1975). In addition to assessing the quantitative importance of background, they have attempted to trace out the extent to which background affects economic standing by affecting cognitive skills and educational attainment. In the process of decomposing the effects of background into direct and indirect components, sociologists have estimated standardized regression coefficients for ability and schooling in models of occupational status and earnings. This work has brought them close afield to interests usually pursued by economists.

Economists of the human capital persuasion have had to contend with the possibility that what appear to be the effects of schooling are, in fact, the effects of the determinants of schooling. Concern with this question has usually centered on the impact of ignoring family background and tested mental ability when estimating the effects of schooling on earnings (Griliches and Mason, 1972; Taubman and Wales, 1974; Welch, 1974).

Both sociologists and economists have usually equated family background with measures of socioeconomic position. Variables which are commonly employed include parental education, family size, and father's occupational status. Critics have been quick to point out that potentially important background measures, such as parental income, are usually omitted (Bowles, 1972).[1] The problem is further complicated by the fact that families may systematically confer advantages and disadvantages in ways that are unrelated to socioeconomic position. "Family climates" and other elusive factors may well vary between families which are equal on all conceivable measures of socioeconomic status and demographic characteristics. If that is true, the explained variance in ordinary models

The research in this paper was supported by the following grants: MDTA Dissertation 91-25-75-43, Ford and Carnegie Foundations to Center for Educational Policy Research, U.S. Department of Labor and National Institute for Education. NIE-G-74 to Center for Educational Policy Research, and Poverty Institute.

of status attainment underestimates the explanatory power of family background. Moreover, if the unmeasured aspects of family background which affect education are correlated with those which affect occupational status or income, controlling measured socioeconomic variables will not suffice to eliminate biases in the education coefficients due to background.

An alternative definition of family background is all those factors, both measured and unmeasured, which produce resemblance on outcomes among siblings. If the effects of family background do not vary systematically by birth order or other within family factors, and if the characteristics of one sibling do not directly affect the characteristics of another, the sibling correlation on an outcome represents the total proportion of variance which background explains.[2] If the entire effect of background defined in this way was produced by measured socioeconomic variables, the R^2's from ordinary individual level regressions would be the same as the sibling correlations.[3] Blau and Duncan (1967) report, however, that this is not the case for educational attainment. My data suggest that it is not the case either for occupational status and earnings. They suggest that ordinary socioeconomic variables are very imperfect measures of family background.[4] Models of the attainment process which ignore this, not only underestimate the overall effects of background, but may also overestimate the extent to which ability and schooling mediate the impact of background on economic attainment.

If the omitted aspects of family background which affect schooling and economic outcomes are uncorrelated, researchers who rely on socioeconomic measures to control background are on safe ground. But if such factors are correlated, estimates of the effects of schooling will be biased to some extent even if socioeconomic background is controlled. By running regressions on sibling differences (or on deviations from pair means) one can control all factors which vary across families and which brothers share. The effects of schooling or other variables such as tested ability measured within families cannot be biased by family background.[5] They can, unhappily, still be biased by unmeasured characteristics which vary between brothers.

Work in status attainment research and in econometric analyses of the effects of education is hampered not only by inadequate measures of family background, but also by the scarcity of data that include ability measures. From the point of view of sociology, which takes a substantive interest in the effects of cognitive skills, the problem is one of scarcity pure and simple. From the point of view of economics, the problem is also conceptual. Traditional cognitive tests may not capture what economists mean by "ability"—i.e., the ability to earn a higher wage irrespective of schooling. Viewed in this light, test scores are possibily error-ridden proxies for "true" ability. However, until economists can specify what such ability is, we will have to be content with the measures which are available. The availability in the Kalamazoo Brothers sample of early test scores for men over 35, adds somewhat to the small stock of existing data which allow useful analysis of the interrelationships among background, ability, schooling, and economic success.[6]

This paper reports the results of my efforts to use the Kalamazoo

Brothers data to estimate the effects of family background on cognitive skills, educational attainment, occupational status, and earnings, and to control family background when estimating the effects of cognitive skills and education on occupational status and earnings. In Section 2, I describe the sample and the variables. In Section 3, I compare the sibling correlations predicted by the effects of measured background to those actually observed, and compare the magnitude of sibling differences to the magnitude of differences between randomly chosen individuals. I also develop alternative models representing the effects of background. And in Section 4, I compare the results of within-pair regressions to individual level regressions. Section 5 summarizes my results and suggests their implications for further research.

Section 2 Sample and Variable Descriptions

The Kalamazoo, Michigan public school system has preserved the results of its standardized testing program since the program's inception in 1928. During the summer of 1973, I selected a sample of males from the records of sixth grade scores for the years 1928 to 1950. I used school census and enrollment records to determine siblingship. This procedure resulted in a potential sample of 2782 individuals from 1224 sets of brothers.

I was able to trace 1612 of the original 2782 individuals in the sample. Of these, 1243 completed a follow-up telephone interview during September 1973 to May 1974; 152 were dead, 52 were never directly contacted and 165 refused to be interviewed. When an interview was conducted with the first brother to be contacted in any set, the respondent was asked to report the schooling, occupation, and earnings of his other brothers who were also in the sample. I concluded that the reports of brothers' occupations and earnings are too unreliable to be substituted for self reports (Olneck, 1976a; Chapter 4), so only men who themselves completed an interview and who could be paired with at least one brother who also completed an interview are included in the present analyses. 916 respondents satisfied that criterion. However, item nonresponse on background variables, initial occupation, and earnings by one or both brothers in a pair lead to further attrition. The analyses reported here are for 692 individual respondents, or 346 weighted pairs.[7] Differences between the means, standard devitations, and correlations for the 1243 men interviewed and the 692 men comprising the present sample are negligible (Olneck, forthcoming, Tables 2 and 11). The average test score for men in this sample is only 3.66 points higher than for men who were not interviewed (i.e., 100.89 v. 97.23). However, comparisons with national and regional data do suggest upward biases on some crucial variables.

Table 1 presents the means and standard deviations for the variables employed in the present analyses. They are compared to means and standard deviations for respondents also aged 35 to 59 from the 1973 replication of the nationally representative "Occupational Changes in a Generation" Survey.[8]

The Kalamazoo respondents are clearly advantaged on parental

Table 1. Means and Standard Deviations of Variables in the Kalamazoo Brothers Sample (N=692) and the 1973 Occupational Changes in a Generation Replication Sample, Men 35 to 59 (N=9398)

Variables		Means		Standard Deviations	
		Kalamazoo	OCG II	Kalamazoo	OCG II
1.	Age	46.13	46.43	6.02	6.94
2.	Test Score	100.89	NA	15.32	NA
3.	Father's Education[a]	9.51	7.90	3.33	3.97
4.	Father's Occupation[a]	38.33	28.29	22.52	21.83
5.	Siblings[a]	3.72	3.83	2.53	2.73
6.	Education	13.20	11.84	2.73	3.29
7.	Initial Occupation	39.51	33.66	23.80	25.18
8.	Current Occupation	49.91	43.18	23.17	25.65
9.	1973 Earnings (Kalamazoo) or Income (OCG II)	16745.66	12821.50 (12775.33)[b]	7633.78	9729.89 (7757.91)[b]
10.	Natural Logarithum of 1973 Earnings (Kalamazoo) Income (OCG II)	9.62	9.19 (9.25)[b]	0.45	1.07 (0.71)[b]

Table 1 Continued (2)

Variable Definitions in the Kalamazoo Sample

1. Age - 1973 minus school record of year of birth.

2. Test Score - Score on Terman group test administered in the 6th grade or score on Otis group test adjusted for scaling differences and trends in parental education, father's occupational status, and family size. See Olneck (forthcoming) for adjustment procedure. Three-quarters of the respondents took the Terman test.

3. Father's Education - Normative years completed (eg., high school graduate is coded 12 even when it took 13 years to finish).

4. Father's Occupation - Duncan Socioeconomic Index. See Duncan (1961).

5. Siblings - Number of siblings who grew up in respondent's family.

6. Education - Normative years completed.

7. Initial Occupation - Duncan score for first full-time civilian job after completion of reported level of schooling.

8. Current Occupation - Duncan score for current job.

9. 1973 Earnings - Expected annual earnings for 1973. Interviewers recorded only the interval in which respondents earnings fell. Reluctant respondents were encouraged to name an interval.

Interval	Coding	Percentage among 1243 interviewees
Under 1000	500	0.2%
1000-1999	1500	0.0
2000-2999	2500	0.1
3000-3999	3500	0.1
4000-4999	4500	0.6
5000-5999	5500	0.4
6000-6999	6500	1.4
7000-7999	7500	1.7
8000-9999	9000	8.8
10000-11999	11000	15.8
12000-13999	13000	17.8
14000-16999	15500	19.4
17000-19999	18500	10.2
20000-24999	22500	11.3
25000 and over	34000	12.1

Notes

a. Errors in these background measures appear random (Olneck, 1977, Chapter 5). Self-reported outcomes correlated as well with background reported by brothers as with self-reported background. Therefore, when reports of father's education or occupation, or number of siblings were missing for a respondent I substituted the report(s) provided by his brother where available. I deleted pairs in which both brothers failed to report a background measure.

b. OCG II income recoded to Kalamazoo coding scheme.

background and adult attainment compared to men of the same age in the nation as a whole. This is due, in part, to characteristics of Kalamazoo. The city has traditionally been an area of skilled employment. It has also had a public college (now university) for some years. The differences between the Kalamazoo and OCG II samples may also be due to my sampling and follow-up procedures. These did not include men who grew up in neighboring farm communities, and they were not likely to result in tracing men whose families left Kalamazoo in the years following the respondent's enrollment in 6th grade unless there were relatives still in Kalamazoo in 1973-1974.[9]

The OCG II sample which I looked at includes proportionately more men 55 to 59 than my sample does. This will tend to exaggerate the differences between the Kalamazoo and national data. Ninety percent of the men I interviewed were between 35 and 54.[10] Among U.S. married men aged 35 to 54 and living with their wives, average 1973 earnings were 15,000 dollars.[11] This is only 1250 dollars less than the average earnings in the Kalamazoo sample. Ninety three percent of my respondents were married and living with their wives.

Table 2 compares correlations in the Kalamazoo sample to those in the OCG II sample. Correlations between measures of attainment are generally similar in the two samples. Differences between correlations involving Ln Earnings and Ln Income are due to differences in coding.[12] The larger correlation between Age and Father's Occupation in the Kalamazoo sample may indicate that younger respondents in that sample come from atypically higher status families. It may, however, indicate that shifts in the occupations held by fathers were more rapid in Kalamazoo than in the nation as a whole.

The most disturbing difference between the correlations from the Kalamazoo and OCG II samples is that Occupation and recoded Income are significantly more highly correlated with Father's Occupation and Father's Education in the OCG II sample. It is tempting to attribute this to the local nature of the Kalamazoo sample.[13] However, the correlations between Education and Initial Occupation on the one hand, and Father's Education and Father's Occupation on the other, are not significantly lower in the Kalamazoo sample than in the OCG II sample. This suggests that the Kalamazoo sample is comprised of respondents whose later but not earlier attainments are usually independent of their parental backgrounds. This, in turn, suggests that rather than the process of attainment being atypically "meritocratic" in Kalamazoo, it is likely there is a success bias in my sample composition.[14] If this is true, the Kalamazoo data would underestimate the impact of measured background characteristics, and would also underestimate biases in the effects of ability and schooling that are due to measured background. Unless the sibling correlations in the Kalamazoo data are lower than those which would be found in national samples, they might correspondingly exaggerate the relative importance of unmeasured background characteristics. There is little evidence that the sibling correlations in the Kalamazoo data are atypically low.

Table 3 presents the correlations between brothers' characteristics for the Kalamazoo sample. Like its predecessor, the 1962 OCG I survey,

Table 2. Correlations Among Variables in the Kalamazoo Brothers Sample (N=692) and the 1973 Occupational Changes in a Generation Replication Sample, Men 35 to 59 (N=9398) OCG shown below.

		1	2	3	4	5	6	7	8	9	10
1.	Age	1.000									
		1.000									
2.	Test Score	-.164	1.000								
		NA	NA								
3.	Father's Education	-.182	.261	1.000							
		-.121	NA	1.000							
4.	Father's Occupation	-.165	.260	.470	1.000						
		-.060*	NA	.501	1.000						
5.	Siblings	.066	-.276	-.250	-.224	1.000					
		.087	NA	-.308	-.295	1.000					
6.	Education	-.184	.576	.400	.383	-.328	1.000				
		-.136	NA	.454	.423	-.357	1.000				
7.	Initial Occupation	-.140	.445	.350	.391	-.256	.716	1.000			
		-.112	NA	.356	.426	-.302	.659*	1.000			
8.	Current Occupation	-.105	.453	.215	.218	-.220	.591	.563	1.000		
		-.067	NA	.340*	.392*	-.282	.624	.630*	1.000		
9.	Earnings (Kalamazoo) or Income[a] (OCG II)	-.071	.359	.171	.212	-.155	.431	.411	.482	1.000	
		-.021	NA	.228	.261	-.191	.388	.378	.453	1.000	
		(-.038)	(NA)	(.260)*	(.298)*	(-.216)	(.452)	(.429)	(.521)	(1.000)	
10.	Ln Earnings (Kalamazoo) or Ln Income[a] (OCG II)	-.083	.360	.160	.197	-.154	.407	.386	.409	.938	1.000
		-.048	NA	.167	.172	-.134	.292*	.256*	.336	.612*	1.000
		(-.058)	(NA)	(.233)	(.243)	(-.188)	(.416)	(.356)	(.466)	(.859)*	(1.000)

a. Correlations in parentheses pertain to OCG II income coded to Kalamazoo coding scheme.

* OCG significantly different from Kalamazooo at the .05 level.

Table 3. Correlations Between Brothers Characteristics (N=346 weighted pairs)

	AGE'	IQ'	ED'	FIRSTOC'	OC'	EARN'	LN EARN'
AGE	.587						
IQ	-.158	.469					
ED	-.157	.400	.549				
FIRSTOC	-.142	.326	.427	.394			
OC	-.120	.300	.378	.321	.309		
EARN	-.032	.178	.285	.231	.225	.237	
LNEARN	-.050	.169	.269	.211	.218	.219	.220

AGE = Age
IQ = Test Score
ED = Education
FIRSTOC = Initial Occupation
OC = Current Occupation
EARN = Earnings
LNEARN = Natural Logarith of Earnings

Primes denote the second member of a given pair. Correlations were computed from a tape on which every pair appears twice, with order reversed. This makes the product moment correlations equal to intraclass correlations.

the 1973 OCG survey asked respondents to report on a brother's educational attainment. Correlations between a respondent's characteristics and his brother's education in the Kalamazoo sample are quite similar to analogous correlations in a subsample of OCG II respondents 35 to 59 who reported their brothers' educations.[15] Sibling correlations on cognitive ability vary depending on the nature, reliability, and timing of the test. My correlations involving Brother's Test Score include no aberrant values.[16] Those involving Brother's Initial Occupation tend to be somewhat higher than analogous correlations reported by Behrman, Taubman and Wales (1976) for fraternal twins, but the differences are not generally large, and in the case of the Initial Occupation-Ln Earnings cross sib correlation there is virtually no difference. My correlations involving Brother's Occupation are similar to those reported elsewhere, with the exception of Behrman, Taubman and Wales (1976), whose value for the correlation between the Duncan scores of DZ twins in the NAS-NRC sample is unusually low.[17] My correlations involving Brother's Earnings are difficult to assess. There are few other studies which have data on brothers' earnings. My correlations tend to lie in the middle of values reported elsewhere. Because of small sample sizes, age restrictions, and unusual sample definitions, these other studies are suspect as regards their generalizability. But that is true also of the Kalamazoo data.[18] This means that my results with respect to the importance of family background on earnings should be viewed with even more caution than my other results.

Section 3 The Impact of Family Background

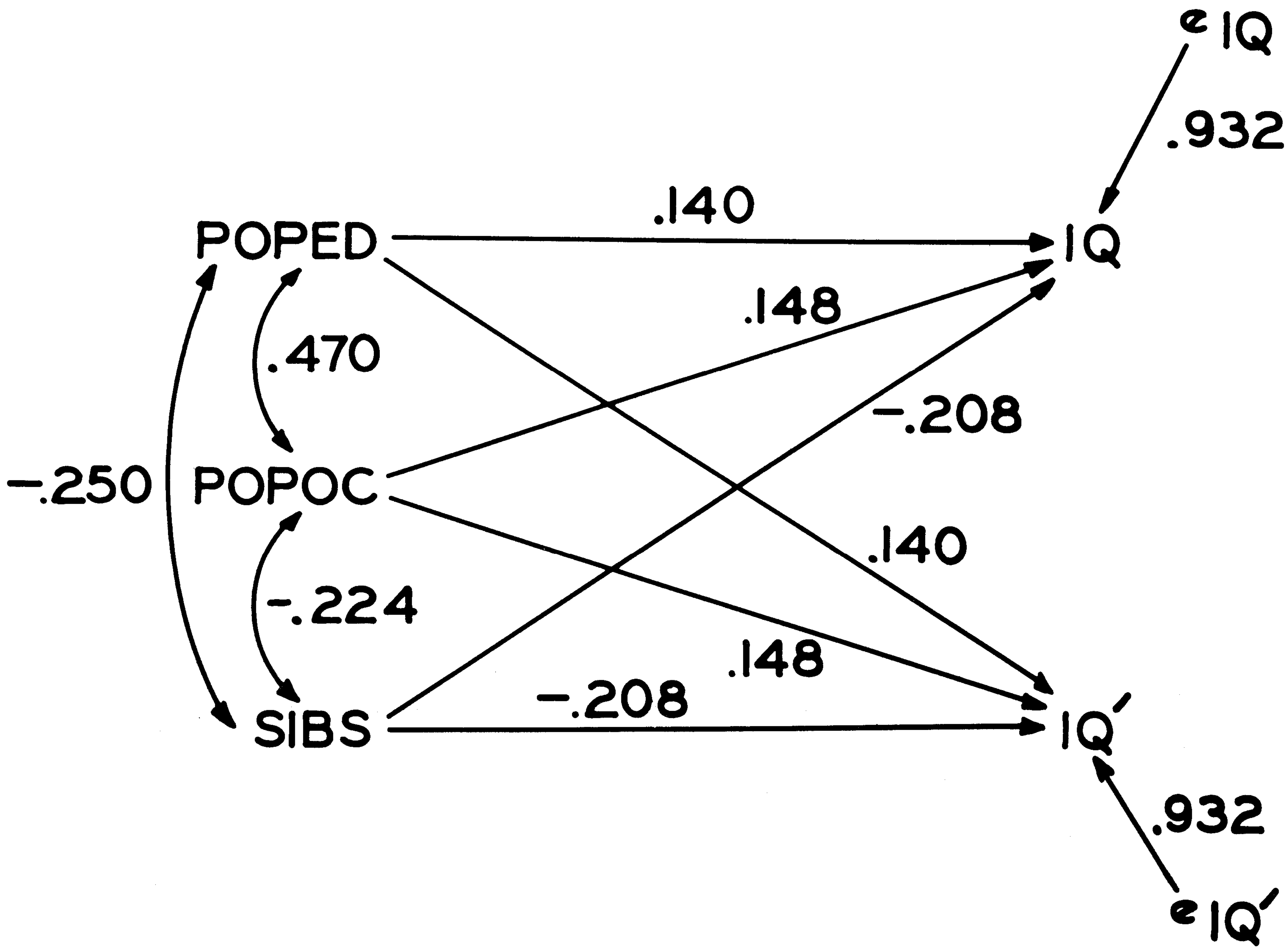
This section considers the overall impact of family background on sons' characteristics, and the directions through which the influences of family background are passed. It does not consider the absolute effects of any given measured background characteristic.[19]

Sibling Resemblance

If family background were adequately measured by socioeconomic variables, if on the average background characteristics affected each brother in a family to the same degree, and if the individual characteristics of one brother did not directly affect the characteristics of another brother, the correlation between brothers on any outcome could be correctly predicted from a path model relating the outcome to background measures. Figure 1 presents such a model, based on the regression of Test Score on Father's Education, Father's Occupation, and Siblings for the 692 individuals comprising my sample.[20] The diagram simply applies the results of the regression to the test scores of two brothers rather than to the score of only one individual.

The fundamental path theorom expresses the correlation between two endogenous variables as $r_{ij} = \Sigma p_{ik} r_{jk}$, where r_{ij} is the correlation being analyzed, p_{ik} is a path (i.e., standardized regression coefficient)

Figure 1: Path Model Relating to Test Scores of Two Brothers to Measures of Socioeconomic Background



from variable k to the second of the two variables (i.e., i), and r_{kj} is the correlation between the first of the variables (i.e., j) and variable k.[21] Apply the path theorom to Figure 1, we can predict the correlation between brothers' test scores from Equation (1).

$$(1) \quad r^*_{IQ',IQ} = p_{IQ',eIQ'}\, r_{eIQ',IQ} + p_{IQ',IQ}\, r_{IQ,IQ} + p_{IQ',POPED}\, r_{POPED,IQ} + p_{IQ',POPOC}\, r_{POPOC,IQ} + p_{IQ,SIBS}\, r_{SIBS,IQ}$$

Since $r_{eIQ',IQ}$ and $p_{IQ',IQ}$ are both assumed to equal 0, rewriting Equation (1) with appropriate values gives Equation (2).

$$(2) \quad r^*_{IQ',IQ} = .140\,(.261) + .148(.260) + (-.208)(-.276) = .132.$$

If the correlation between brothers' test scores arises only because of the effects of Father's Occupation, Father's Education, and Siblings, we would expect the sibling correlation on test scores to be 0.132. This is exactly the proportion of variance in individual score explained by the regression of Test Score on the three background measures. This can be seen by comparing the equation predicting the sibling correlation to the equation for R^2 for a dependent variable, controlling one or more independent variables.

The equation for R^2 is $R^2_{i.kj} = \sum_{k}^{j} p_{ij} r_{ij}$, where $R^2_{i.kj}$ is the proportion of variance in i explained by the regression of i on variable k and j, p_{ik} is the path from k to i, and r_{ik} is the correlation between i and k. Since the correlation between measured background variables and individual outcomes is assumed to be the same for all brothers (e.g., $r_{IQ',POPED} = r_{IQ,POPED}$), Equation 2 is nothing more than the equation for R^2 in a regression of Test Score on Father's Education, Father's Occupation, and Siblings.

Column 1 of Table 4 gives the predicted sibling correlations for test scores, educational attainment, initial occupation, current occupation, earnings, and ln earnings. Column 2 gives the observed correlations. The results in Table 4 show that analyses which equate family background with measured socioeconomic variables will fall far short of accounting for resemblance among brothers on test scores, education, and economic attainment. Moreover, even if the actual value for the sibling correlation on test scores is assumed prior to predicting other sibling correlations, and test scores are incorporated into models predicting subsequent outcomes, the predictions will fall short. There are substantial advantages and disadvantages associated with family to family variations within equal levels of measured socioeconomic background, and which are not mediated by tested ability.[22]

Unless the brothers in the Kalamazoo sample are unusually similar,

Table 4. Comparison of Sibling Resemblance Predicted by the Effects of Socioeconomic Background to Observed Sibling Resemblance (N=346 weighted pairs)

Variable	Predicted Sibling Correlation[a]	Observed Sibling Correlation	Residual Standard Deviation Controlling Socioeconomic Background[b]	Residual Standard Deviation Controlling Brothers' Shared Background[c]
1. Test Score	.132	.469	14.27	11.16
2. Education	.253	.549	2.36	1.83
3. Initial Occupation	.209	.394	21.17	18.53
4. Current Occupation	.088	.309	22.13	19.26
5. Earnings	.061	.237	7397.29	6668.10
6. Ln Earnings	.055	.220	0.44	0.40

a. R^2 from regressions in which Father's Education, Father's Occupation, and Siblings are the independent variables.

b. Father's Education, Father's Occupation, Siblings.

c. Calculated as $[1-r_{sib}]^{1/2}$ S, where r_{sib} is the sibling correlation and S is the standard deviation of the dependent variable reported in Table 1. This is <u>not</u> the observed within-pair standard deviation $[(1-r_{sib})/2]^{1/2}$. The observed within-pair standard deviation is less than the total standard deviation even when the sibling correlation is zero.

it is unlikely that I have substantially overestimated the relative importance of unmeasured aspects of family background for any outcome with the exception of current occupational status. Except for Current Occupation, R^2 from analogous regressions for the OCG II sample aged 35 to 59 are quite similar to those for my sample.[23] For current occupation, R^2 is appreciably higher in the OCG II data than in the Kalamazoo data. This suggests that unmeasured background factors may not be as important for that outcome as my data suggest, unless, of course, the sibling correlation on occupation in the nation as a whole is much larger than it is in my data.

Nor is it likely that I have overestimated the importance of unmeasured background factors relative to measured factors because of measurement error. When I attempt to correct my correlations for measurement error, R^2's rise, but so do sibling correlations. Predicted sibling correlations based on corrected data underestimate the corrected sibling correlations by almost the same proportions as in the observed data. The only outcome for which there is appreciable improvement in prediction is initial occupation.[24]

Differences Between Siblings

If the distributions of the outcome measures were normal, we could calculate the average differences between two randomly picked individuals and compare them to the average differences between two randomly chosen brothers.[25] Because the distributions of outcome variables depart to some extent from normality, we must calculate average differences between brothers directly, and, assuming similar distributions within and between pairs, infer the average differences between randomly picked individuals from the observed differences between brothers and the sibling correlations.

The average pair of brothers in the Kalamazoo sample differs by around 12 points on test scores, 1.78 years on educational attainment, 19 points in initial occupational status, 21 points on current occupational status, 6690 dollars on earnings, and 0.406 on ln earnings. Assuming that the ratio of differences between randomly chosen individuals and pairs of brothers is $1 : (1 - r_{sib})^{1/2}$, suggests that the average difference between randomly paired individuals in my sample is 16 points on test scores, 2.66 on years of schooling, 24 points on initial occupational status, 25 points on current occupational status, 7690 dollars on earnings, and 0.460 on ln earnings.

Thus despite the results in Table 4 showing that family background has substantially larger effects than ordinary sociological analyses might imply, the effects are nonetheless modest when viewed against the overall degree of inequality in outcomes. This is especially true of earnings. The average difference between brothers on earnings is 87 percent as large as the difference between random individuals. Eliminating earnings differences among men raised in the same home would do far more to reduce variance in income than would eliminating differences between men raised in different families. If brothers earned the same amount as one another, while family to family differences in earnings remained unaltered, the standard deviation of the resulting distribution of earnings in the Kalamazoo Brothers sample would be 3716/7634 = 48.7 percent of the present standard deviation. But if differences explained by family background were eliminated, while differences among brothers were unaltered, the resulting standard deviation of earnings would be 6668/7634 = 87.3 percent as large as the present standard deviation.[26]

A Note on Spacing

If families treat brothers who are close in age more alike than they treat brothers who are far apart in age, or if brothers who are close in age encounter more common influences outside the home than do widely-spaced brothers, we would expect brothers who are far apart in age to resemble each other less than closely-spaced brothers. On the other hand, if sibling resemblance is due to genetic influences, or if the extent to which brothers have similar environments does not depend on how close in age they are, we would expect the degree of resemblance between brothers to be unaffected by age differences. My evidence is not fully consistent with either alternative. It generally supports the latter conclusion.

Absolute differences on all outcomes except current occupational status are unrelated to age differences. The correlation between absolute age difference and absolute status difference is 0.145 (t=2.70). Occupational differences between brothers do not systematically favor older or younger members of a pair. The effect of age differences on occupational differences among brothers is insignificant. Therefore while brothers who are far apart in age are likely to differ more from one another on occupational status than brothers who are close in age, the direction of the difference cannot be predicted.

If the overall variances of variables were different among individuals who came from widely-spaced pairs than they are among individuals from closely-spaced pairs, than they are among individuals from closely-spaced pairs, sibling correlations could differ even though absolute differences did not vary by age-spacing. To investigate this possibility, I divided my sample into pairs of brothers 3 or less years apart in age, and pairs more than three years apart in age. Table 5 shows the sibling correlations and the within-pair standard deviations for the two groups.

The only dramatic difference between the results for the two groups involves current occupation. The correlation between brothers' occupations is 0.469 among pairs 3 or less years apart in age, but only 0.181 among men more than 3 years apart in age. Since brothers from closely-spaced pairs are not significantly more likely to have similar educational attainments and initial occupations than brothers from widely-spaced pairs, this result is puzzling. It suggests that common family background has a direct impact on occupational status for closely-spaced brothers, but that widely-spaced brothers resemble each other on occupational status only to the extent that they have similar amounts of education and hold similar jobs when they finish school.[27] If this explanation were correct, however, I would expect a similar result with respect to earnings. No such

Table 5. Sibling Correlations and Within-pair Standard Deviations for Brothers Three or Less Years Apart in Age and for for Brothers More Than Three Years Apart in Age

Variable	Sibling Correlation		Within-Pair StandardDeviation	
	3 or Less (N=155) pairs	More than 3 (N=197) pairs	3 or Less (N=155) pairs	More than 3 (N=197) pairs
Test Score	.516	.434	7.47	8.24
Education	.570	.531	1.32	1.27
Initial Occupation	.424	.379	13.06	12.97
Current Occupation	.469	.181*	12.02	14.75*
Earnings	.266	.183	5005	4542
Ln Earnings	.196	.201	.331	.261*

*Significantly different at the .05 level.

pattern is apparent in these analyses, so in the absence of further evidence, it seems reasonable to attribute the finding concerning occupation to sampling error,[28] and to conclude that the extent to which brothers enjoy common background influences is similar regardless of age differences. The data cannot be used to determine if this is because family-related environmental influences are stable, or if it is because sibling resemblance on outcomes is due to genetic resemblance between brothers.

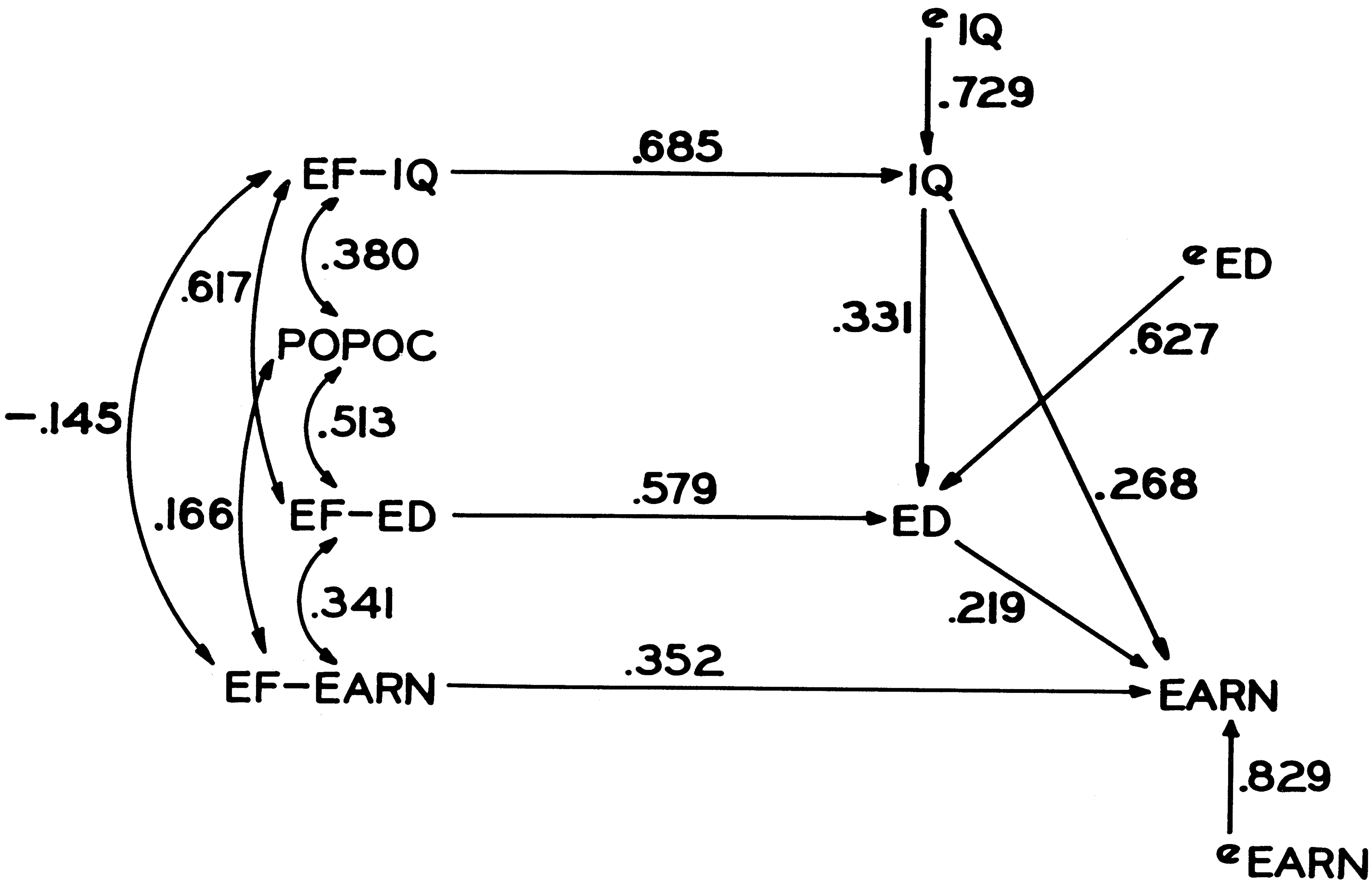
Models of the Effects of Family Background

In order to investigate the extent to which family background exercises direct effects on outcomes, the extent to which families that confer advantage on one outcome do so on others, and the extent to which the effects of schooling and ability transmit background rather than introduce variation in outcomes that is independent of background, I constructed two models which account for the observed individual and cross-sibling correlations among test scores, education, and earnings.[29] They are shown in Figures 2 and 3.

In Figure 2, the effects of family background are represented as deriving from a set of correlated, but unmeasured variables which affect one and only one outcome. The values of the paths from these variables represent the effects of family background necessary to account for observed sibling correlations. The hypothetical variables themselves (except EF_{IQ}) may be thought of as representing advantages or disadvantages family membership confers net of the effects of measured variables. The variable EF_{EARN}, for example, measures the tendency of two brothers to have similar deviations from the earnings expected for each of them on the basis of educational attainment and test scores. EF-IQ represents the total effects of shared background on brothers' test scores. The correlations among the hypothetical variables measure the extent to which families confer similar net advantages or disadvantages across outcomes. Father's Occupation is included in the model, but constrained to have no effect. It is included to suggest the relationships between variables measuring the overall impact of family background and more traditional measures of socioeconomic status.[30]

Figure 2 suggests that the effects of family background on years of education are not explained by sibling resemblance on tests scores (See f.n. 22). $(.579)^2/.549 = 61.1$ percent of the correlation between brothers on education arises as the result of background effects which are not mediated by or shared with the effects of test scores. $(.352)^2/.237 = 52.3$ percent of the correlation between brothers' earnings is independent of the effects of background on test scores and education. The data do not enable us to determine what it is that brothers share which accounts for the continuing effects of background on education and earnings. Corcoran, Jencks, and Olneck (1976) suggest that the weak correlation between a hypothetical variable determining earnings, and Father's Occupation, evident in several data sets, argues against such a variable representing economically productive skills. (Note $r_{EF\text{-}EARN,POPOC}$ =

Figure 2: Model of Individual
Attainment Omitting Occupational Status



.166 in Figure 2 below.) They suggest that it instead may proxy shared preferences for pecuniary versus nonpecuniary rewards. It is possible, however, that the variable represents a combination of personality characteristics, unmeasured skills, values and shared information which bear varying relationships to Father's Occupation. Attempts to reject or establish unitary definitions of such a variable are, therefore, potentially misleading.[31]

The correlation among the hypothetical variables indicate that families who have sons with higher test scores also tend to have sons whose educational attainments exceed the attainments expected on the basis of test scores alone, but that net earnings advantages associated with family membership are not strongly related to net educational advantages or to overall test score advantages. Indeed, families whose sons have test scores above the mean tend, albeit weakly, to have sons whose earnings are below the earnings expected on the basis of test scores and education alone. (Note in Figure 2 that while $r_{EF-ED,EF-IQ}$ equals 0.617, $r_{EF-EARN,EF-ED}$ equals only 0.341, and $r_{EF-EARN, EF-IQ}$ equals -0.145.)

Figure 3 presents a model in which the overall, rather than the net effects of family background on individual outcomes are represented. The effect of each hypothetical variable is simply the square root of the sibling correlation for the outcome associated with the variable. The correlations among the hypothetical variables are calculated by using cross-sib correlations (e.g., $r_{ED,IQ'}$), and measure the tendency of brothers who share advantages on one outcome to share advantages on others. The error terms in the model are the square root of the variance not explained by family background. The correlations between an individual characteristic (e.g., $r_{ED,IQ}$) are accounted for by the effects of family background and a correlation between error terms. For example, the correlation between earnings and education is expressed in Equations 3 and 4.

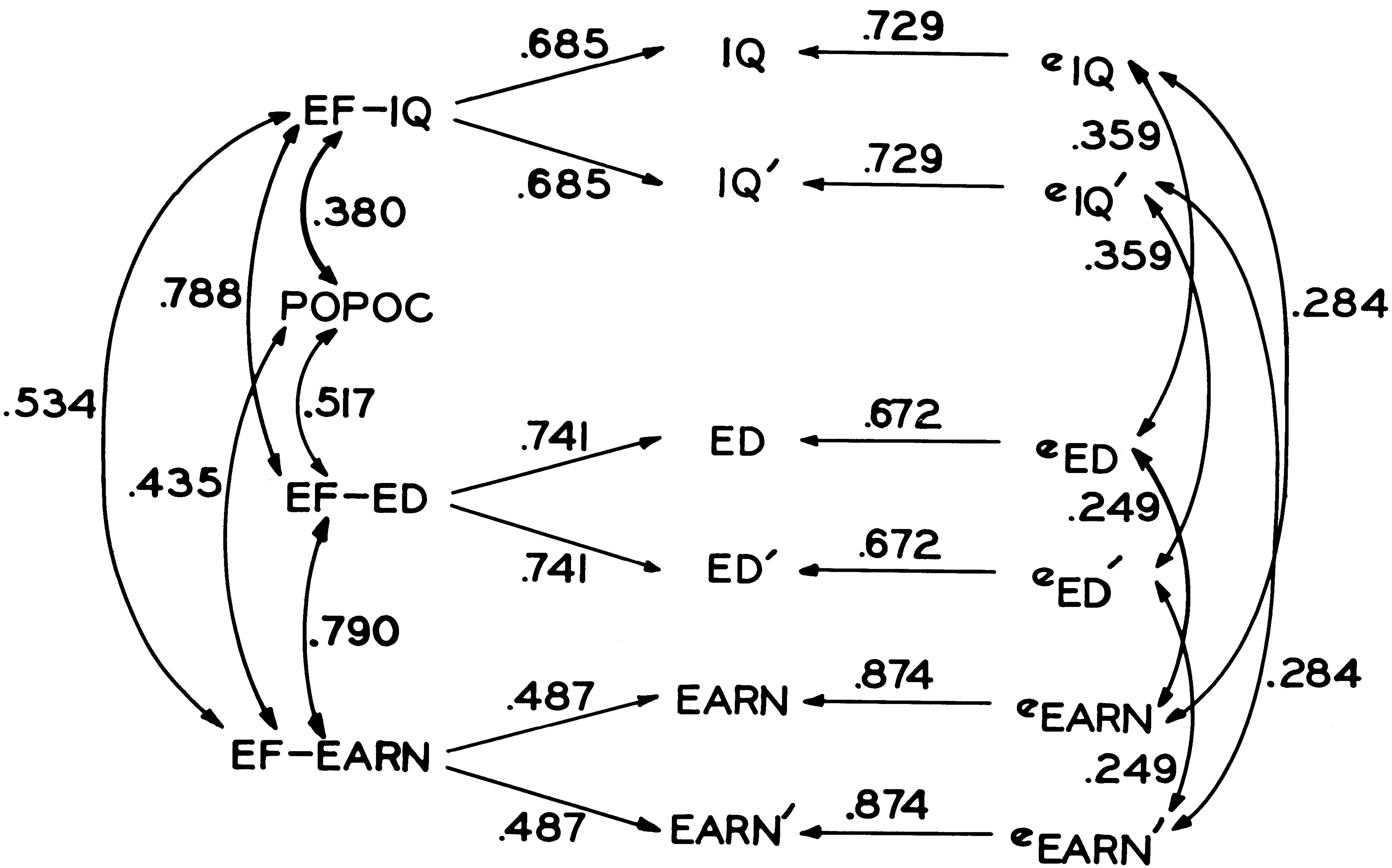
$$(3) \quad r_{EARN,ED} = p_{EARN,EF-EARN} r_{EF-EARN,ED} + p_{EARN,e-EARN} r_{e-EARN,ED}$$

$$(4) \quad .431 = .487\,(.790)\,(.741) + .874\,(.249)\,(.672)$$

The model shown in Figure 3 allows us to determine the extent to which brothers who are advantaged on one outcome tend to have similar shared advantages on other outcomes, and to determine the extent to which individual level effects are independent of family background.

The inter-correlations among the hypothetical variables in Figure 3 suggest that brothers who come from families that are unusually effective in conferring educational advantages, also tend strongly to come from families that are unusually effective in their influence on both test scores and earnings, but families that are unusually effective in their influence on test scores are not as likely to be similarly effective in their influence on earnings.

Figure 3: Model Representing the Overall Impact of Family Background on Test Scores, Education and Earnings (Prime denotes brother.)



Sociologists have sometimes attempted to use the results from models like that shown in Figure 2 or models which incorporate only measured background variables, to estimate how much of the variance in outcomes such as occupational status or earnings is due solely to the independent effects of cognitive skills or education (See especially Duncan, 1968). These estimates are calculated by squaring standardized regression coefficients. Such attempts are potentially misleading because they may confuse different meanings of independence.

The effects of the endogenous variables in Figure 2 are independent of family background in that their values were calculated by holding background constant. They are free from the biasing effects of family background factors common to outcomes and their determinants. But the path coefficients of endogenous variables in Figure 2 are equal to the unstandardized regression coefficients of within-pair regressions multiplied by the ratios of the <u>total</u> standard deviations of independent and dependent variables. They, therefore, do not represent effects that produce variance in outcomes that is entirely orthogonal to family background. Effects whose magnitudes are independent of family background may nevertheless contribute to intergenerational status inheritance and sibling resemblance.[32]

I have used the results shown in Figure 3 to determine the extent to which the correlations among Test Scores, Education, and Earnings involve familial and nonfamilial components. Equations 5 and 6 represent the correlation between Test Scores and Education as the sum of a family related component and a component arising only from the association between scores and attainment within families.

$$(5)\quad r_{ED,IQ} = p_{ED,EF\text{-}ED}\, r_{EF\text{-}ED,EF\text{-}IQ}\, p_{IQ,EF\text{-}IQ} + p_{ED,e_{ED}}\, r_{e_{ED}e_{IQ}}\, p_{IQ,e_{IQ}}$$

$$(6)\quad .576 = (.741)(.788)(.685) + (.672)(.359)(.729)$$

$$.576 = .400 + .176$$

The results in Equation 6 show that .400/.576 = 69.4 percent of the correlation between Test Scores and Education arises because of the association between them across families, and only .176/.576 = 30.6 percent arises because of the within-family correlation between scores and attainment. Similarly, .181/.359 = 50.4 percent of the correlation between test scores and earnings arises because of the within-family correlation between them, and only .146/.431 = 33.9 of the education-earnings correlations is due to the within-family correlation.

These results strongly suggest that relationships generally thought to represent meritocratic processes serve in larger measure to transmit family background, broadly defined, than they do to sever the ties between background and adult status. This may not be disturbing to those for whom meritocratic ideology stresses the mechanisms rather than the results of status allocation, or for those who equate background solely with socioeconomic status, but it should give pause to those for whom so-called merit (or achievement) and equal opportunity are closely linked in principle. Moreover, cognizance of the nonequalizing effects of measured cognitive skills and education should prompt reexamination of our definitions and standards of merit. Those standards might survive reexamination as to their necessity and fairness. I suspect, however, that their appeal lies to some extent in their presumed impact on diminishing the effects of family background, and that presumption is called into question by my results.

Section 4 Controlling Family Background

In order to determine the extent to which unmeasured background factors impart bias to estimates of the effects of cognitive skills and schooling, I ran regressions on sibling differences as well as on individual level data. Table 6 gives the results of these analyses.

Among individuals, a 10 point test score difference is associated with a 1.03 year difference in educational attainment. Controlling measured background variables reduces this effect to 0.81 years, and controlling unmeasured shared background as well reduces it further to 0.59 years. This results suggests that 1 - .59/1.02 = 43 percent of the relationship between test scores and education arises because men with higher test scores tend to come from families which somehow promote educational attainment independently of their sons' abilities. However, this result could also arise if the abilities which vary across families and those which vary within families were different. A single ability measure is insensitive to this possibility. If abilities which vary between families strongly affect education and those which vary within families do not, reduced coefficients for a single ability measure would result when family background was controlled, even though this would not mean that background rather than ability causes higher educational attainment. It would only mean that the effects of the two could not be distinguished without direct measure of multiple abilities.

The results in Equations 4 to 9 indicate that estimates of the effects of cognitive skills on initial occupational status are quite sensitive to controls for family background, but that the effects of educational attainment are robust. Moreover, controlling test score differences among brothers barely reduces the schooling coefficient below the coefficient controlling only brothers' common background. These results suggest that when employers favor better-schooled young men they are either seeking characteristics that are relatively unrelated to cognitive ability and family background, or that they are poor judges of ability and background, and rely on educational credentials as an imperfect guide.[33]

Equations 13 to 18 suggest that controlling measured socioeconomic background is inadequate to eliminate biases in estimates of the effects of test scores and schooling on current occupational status. The coefficient for test scores controlling only measured background is .601/.685 = 87.7 percent as large as the uncontrolled coefficient, while the within-pair coefficient is only .436/.685 = 63.6 percent as large as the uncontrolled coefficient. Similarly, controlling measured background does not reduce the coefficient of education at all, but the within-pair education coefficient is reduced by 1 - (4.002/5.016) = 20.2 percent.[34]

Table 6. Effects of Test Scores and Education (Standard errors of regression coefficients in parentheses; Bracketed coefficients less than 1.96 times their standard errors.)

Dependent Variable	Test Score	Education	$\bar{R}^{2}$ [a]	Residual Standard Deviation	Other Variables Controlled
1. Education	.103 (.006)		.333	2.23	None
2. Education	.081 (.006)		.431	2.06	Socioeconomic Background[d]
3. Δ [b]Education	.059 (.008)		.608	1.71[c]	Brothers' Common Background
4. Initial Occupation	.691 (.053)		.197	21.33	None
5. Initial Occupation	.510 (.053)		.299	19.93	Socioeconomic Background
6. Initial Occupation	.350 (.087)		.420	18.13[c]	Brothers' Common Background
7. Initial Occupation		6.242 (.232)	.512	16.63	None
8. Initial Occupation		5.170 (2.64)	.525	16.40	Socioeconomic Background
9. ΔInitial Occupation		5.576 (.454)	.577	15.47[c]	Brothers' Common Background
10. Initial Occupation	[.076] (.050)	5.997 (.283)	.513	16.61	None
11. Initial Occupation	[.062] (.050)	5.520 (.303)	.525	16.40	Socioeconomic Background
12. Δ Initial Occupation	[.022] (.080)	5.526 (.488)	.576	15.49[c]	Brothers' Common Background
13. Occupation	.685 (.051)		.202	20.70	None
14. Occupation	.601 (.055)		.217	20.50	Socioeconomic Background
15. Δ Occupation	.436 (.090)		.351	18.66[c]	Brothers' Common Background

Dependent Variable	Test Score	Education	$\bar{R}^2$ [a]	Residual Standard Deviation	Other Variable Controlled
16. Occupation		5.016 (.261)	.349	18.70	None
17. Occupation		5.031 (.302)	.347	18.72	Socioeconomic Background
18. Δ Occupation		4.002 (.524)	.407	17.84[c]	Brothers' Common Background
19. Occupation	.255 (.056)	4.192 (.314)	.367	18.44	None
20. Occupation	.254 (.057)	4.280 (.342)	.362	18.50	Socioeconomic Background
21. Δ Occupation	.229 (.092)	3.499 (.557)	.416	17.70[c]	Brothers' Common Background
14. Δ Occupation	.224 (.090)	2.150 (.639)	.441	17.32[c]	Brothers' Common Background, Initial Occupation
15. Earnings	179 (18)		.128	7130	None
17. Earnings	156 (19)		.141	7075	Socioeconomic Background
18. Δ Earnings	170 (31)		.296	6404[c]	None
19. Earnings		1205 (96)	.185	6893	None
20. Earnings		1157 (111)	.184	6895	Socioeconomic Background
21. Δ Earnings		906 (190)	.282	6469[c]	Brothers' Common Background
22. Earnings	83	938 (21)	.202 (116)	6820	None
23. Earnings	82 (21)	914 (126)	.202	6820	Socioeconomic Background

Dependent Variable	Test Score	Education	$\bar{R}^2$ [a]	Residual Standard Deviation	Other Controlled Controlled
24. Δ Earnings	133 (33)	612 (199)	.313	6327[c]	Brothers' Common Background
25. Δ Earnings	111 (32)	[276] (203)	.361	6102[c]	Brothers' Common Background, Occupation
26. Ln Earnings	.0106 (.0010)		.129	.420	None
27. Ln Earnings	.0094 (.0011)		.137	.418	Socioeconomic Background
28. Δ Ln Earnings	.0105 (.0018)		.294	.378[c]	Brothers' Common Background
29. Ln Earnings		.0671 (.0057)	.166	.411	None
30. Ln Earnings		.0642 (.0066)	.166	.411	Socioeconomic Background
31. Δ Ln Earnings		.0499 (.0113)	.268	.385[c]	Brothers' Common Background
32. Ln Earnings	.0055 (.0012)	.0492 (.0069)	.186	.406	None
33. Ln Earnings	.0055 (.0012)	.0480 (.0075)	.186	.406	Socioeconomic Background
34. Δ Ln Earnings	.0086 (.0019)	.0310 (.0118)	.306	.375[c]	Brothers' Common Background
35. Δ Ln Earnings	.0072 (.0019)	[.0094] (.0119)	.364	.359[c]	Brothers' Common Background, Occupation

a. Calculated as 1 - (Error Variance/Total Variance) for individuals.

b. Δ indicates variables defined as sibling differences.

c. Within pair standard deviation corrected for degrees of freedom. Calculated as .5(1.4144)=.707 times the observed standard deviation of residuals for regressions of sibling differences.

d. Father's Education, Father's Occupation, Siblings.

Equation 21 indicates that controlling brother's test score differences reduces the within-pair coefficient of education. The combined ability-family background bias in the occupation-education relationship is 1 - (3.499/5.016) = 30.2 percent. This is larger than the proportionate bias suggested by other data sets which include ability measures.[35]

Equation 26 indicates that a 10 point difference in test scores is associated with an 11.2 percent difference in earnings among individuals.[36] Controlling measured socioeconomic background reduces the effect slightly but among brothers the effect is virtually the same as it is without family background controlled. Moreover, among brothers, the regression coefficient for test score differences controlling schooling and occupation differences is 0.0072. The analogous coefficient for individuals, controlling socioeconomic background, schooling, and occupation differences is only 0.0037 (Olneck, 1976a).

There are three possible explanations for this result. One is sampling error. Crouse (forthcoming) reports that for the Project Talent sibling subsample, the within-pair test score coefficient for ln earnings is lower than the uncontrolled coefficient. Another is that the unmeasured aspects of family background which affect earnings, net of the effects of cognitive skills, are negatively correlated with the unmeasured aspect of background that affect test scores. Family background is consequently a suppressor variable. Figure 2 embodies this interpretation. Finally, standardized tests may measure multiple abilities, some of which exercise large direct effects on earnings and others which may not. Brothers may resemble each other strongly on the abilities related to test scores which have weak direct effects on earnings, but may vary among themselves on the abilities related to test scores which have strong direct effects. This last possibility cannot be tested without direct measures of different abilities or skills.[37]

In the Kalamazoo data, measured socioeconomic background does not bias estimates of the effects of education on ln earnings, but unmeasured aspects of family background do. While the coefficient of education controlling Father's Education, Father's Occupation, and Siblings is virtually identical to the uncontrolled coefficient, the effect of a one year difference in education among brothers is only .0499/.0671 = 74.4 percent as large as the uncontrolled effect.[38]

Equation 34 indicates that the combined ability-background bias in the education-ln earnings relationship is quite large. The within-pair coefficient of education, controlling brothers' test score differences is 0.0310. This is only .0310/.0671 = 46.2 percent as large as the uncontrolled coefficient. My results, along with those of Behrman, Taubman, and Wales (1976), suggest that when researchers work with young samples in which ability differences have small effects, or with samples that control only measured background, they will erroneously conclude that the bias in the education-income relationship is small.[39]

Note on Measurement Error

This chapter has emphasized family background and tested ability as a source of upward biases in the observed effects of schooling. I ignored the likelihood that the effects of schooling are biased downward to some extent because of measurement error. Bishop (1976) has noted that the use of sibling data can exacerbate the problem of measurement error, and has argued that the within-DZ twin pair unstandardized coefficient of schooling in an earnings equation is at a maximum only 83 percent of the true effect. However, the accuracy of educational reports in the Kalamazoo data appears to be slightly higher than in the CPS data Bishop analyzed. My results would indicate that if there were no other omitted variables, the observed within-pair coefficient of education could be 89 percent of the true coefficient.[40]

However, the Kalamazoo data also include an ability measure. The remaining bias in the within-pair education coefficient due to measurement error depends on the relative extent of error in the sibling differences of schooling and test scores. Since the ratio of error variance to the variance of sibling differences in education appears to be smaller than the analogous ratio for test scores, adding test score differences reduces the remaining downward bias in the within-pair education coefficient.[41] Therefore, unless there are important remaining omitted variables, the observed within-pair education coefficient, controlling test score differences, is probably close to 90 percent of the true coefficient. If this were true, the bias in the observed coefficient would still be 1 - (1.11)(.0310)/.671 = 48.7 percent.

Section 5 Summary and Discussion

Standard sociological variables do not adequately measure family background. We would reach this conclusion even if we included measures of parental income, and if we measured background variables more accurately. Family background exercises continuing effects on adult earnings and occupational status that are not mediated by measured ability or educational attainment. Nevertheless, the differences between brothers on measures of economic success are quite large relative to differences in the general population.

Measured ability and education, which are often thought to represent "meritocratic" characteristics, in part because they are presumed to significantly diminish the ties between background and attainment, transmit background more than they reduce its effects. If the correlations between test scores and education, and education and ln earnings arose soley from effects that were orthogonal to background they would be only one-third of their present magnitudes.

Controlling measured socioeconomic variables does not fully eliminate biases due to background in the effects of test scores on educational attainment, and in the effects of education on current occupational status and earnings. The effect of measured ability on earnings among brothers is, however, the same as it is among unrelated individuals. This result is anomolous, and may well be due to sampling error.

My results should encourage sociologists to investigate in more detail the processes by which families influence the destinies of their children. Unitary conceptions of family background do not account for the continuing effects of background on various outcomes. The sources of the net effects of background on one outcome are weakly related to the sources of the net effects of background on other outcomes. (See Figure 2).

They should also encourage econometricians analyzing bias in the income-schooling relationship to posit multiple omitted variables. Family background and test scores both impart bias to the schooling coefficient. While background and test scores might be imperfect measures of a single ability, it is likely that persistence in school and higher earnings are related to more than one common factor.

I would hope, however, that the principal impact of my results would be to encourage others to reconsider the theoretical and ideological underpinnings of research such as mine. Researchers investigating the relationships among family background, test scores, education, and economic success are implicitly engaged in normative discourse even if they only report technical analyses. Their work is part of an ongoing discussion about equal opportunity, and embodies societal commitments to shared conceptions of merit and entitlement. It also embodies assumptions about the importance of individual characteristics for explaining individual attainments.

The choice of test scores and education as explanatory variables is intimately tied to our view that "ability" and "effort" rather than inherited advantage should predominate in the process of economic attainment. Despite the fact that our research only measures the extent to which test scores and education are related to economic outcomes, and does not directly examine the processes by which those relationships arise, we rarely question the identification of IQ and schooling with merit. Until we know more about why better-schooled and higher-scoring individuals are economically favored, we cannot know whose needs and interests are served by the use of so-called meritocratic criteria.

Models of individual attainment embody the assumption that differences in adult success can be explained by differences in individual characteristics. Inquiry centers on whether the important characteristics are those which are "fair" (eg. schooling), "unfair" (eg. background), or unmeasured (i.e. the residual). Two prior assumptions are implicit in the assumption that individual attainments can be explained. One is that the distribution of attainments or rewards is causally produced by the characteristics of individuals. Economists assert this assumption by calling their models "structural equations." The other, which is a corollary of the first, is that the distribution of rewards is not fixed, but will respond to changes in the distribution of individual characteristics.

These assumptions, while normatively appealing because of their affinity with traditional American values about hard work and individual effort, obscure the capriciousness and randomness which my results and those of others suggest characterize the economic game. My results suggest that differences in family background, measured cognitive ability, and schooling are not primary sources of economic inequality among adults. Seventy percent of the variance in earnings in my sample remains unexplained after the effects of background, test scores, and schooling are taken into account.[43] This suggests that research paradigms which inherently reinforce the view that our own economic fates and the overall distribution of economic rewards are generated by personal characteristics should be seriously questioned, and emphasis in economic research should be concentrated on the systemic factors determining inequalities in economic rewards. In sociology, a more fruitful pursuit than the further refinement of path models would be an assessment of the ideological antecedents and impact of the dominance of the status attainment school.

FOOTNOTES

[1]What direct evidence there is suggests that the inclusion of parental income reduces the coefficients of other background measures, but that it does not significantly enhance the explanatory power of measured background. I reanalyzed Sewell and Hauser's sample of 1957 Wisconsin high school seniors, and found that the addition of average parental income from 1957 to 1960 to equations already including Father's Education, Mother's Education, and Father's Occupation did not significantly reduce the residual standard errors for educational attainment, 1964 occupational status, and 1967 earnings.

[2]If the assumptions do not hold, the sibling correlation still reflects the extent to which between family variance exceeds within-family variance, but the interpretation of the correlation becomes ambiguous. If the effects of background vary by birth order, the proportion of variance due to family and to such an interaction could be higher than the sibling correlation. If brothers' characteristics directly affect one another, the sibling correlation exceeds the variance attributable to shared background characteristics. Fortunately, the assumptions that background effects are symmetric by birth order and that interbrother effects are unlikely appear tenable for the Kalamazoo data. See Olneck (1976a, Chapter 4).

Two other caveats are in order. If background factors have different effects for men with no brothers, estimates of explained variance based on sibling data may be misleading for the general population. This possibility cannot be tested for unmeasured background factors. Nor am I familiar with analyses of national data which relate outcomes to measured variables separately for men with brothers and men with no brothers. Such analyses could be conducted with the 1962 and 1973 OCG data (Blau and Duncan, 1967; Featherman and Hauser, 1975).

My definition of "background" includes the effects of genes, but only to the extent that brothers' genetic makeups are correlated. If genes are viewed as an "inheritance", I have underestimated the effects of background even when using sibling data. However, unshared, unmeasured environmental factors whose effects I cannot analyze may also be related to family background in a narrow sense, and in a wider sense are almost definitionally related to background. No methodology can analytically distinguish unmeasured individual "background" factors from "later" influences.

[3]See Equations 1 and 2.

[4]This would be true even if socioeconomic variables were measured without error. While R^2's from equations using corrected variables are

higher than those from equations using observed measures, corrected sibling correlations are also higher. See Olneck (1976a, Table 4.7).

[5] For an early anticipation of this strategy, see Gorseline (1932). For reanalyses of Gorseline's data, see Chamberlain and Griliches (1974). Behrman, Taubman, and Wales (1976) also report within-pair regression results.

The strategy involves two hazards which I discuss in more detail below. The first is that it assumes variables measure the same things within and between families. The second is that it exacerbates biases due to measurement errors (Bishop, 1976).

[6] The Wisconsin 1957 high school seniors (Sewell and Hauser, 1975) are only now in their mid-thirties, and the sample excludes high school dropouts. Published analyses of this sample cover earnings only 10 years after high school graduation. The Project Talent respondents (Crouse, forthcoming$_b$) were only around 28 years old when last surveyed. The effects of cognitive skills on earnings appear to be lower in the early career than later on. [See Hause's (1972) report of Roger's data; Also see Jencks (forthcoming) and Fägerlind (1975).] Unpublished data from the Wisconsin sample also show this effect. This means that analysts who have relied on younger samples may have prematurely concluded that the ability bias in the income-schooling relationship is small. For example, Griliches and Mason (1972) concludes that the bias in post-military schooling in the NORC Veterans sample is only 10 percent. I found the bias in the coefficient for total schooling for respondents 30-34 in that sample to be 42 percent. See Olneck (1976b).

Unfortunately, samples of older men which include test scores are rare, and, invariably, flawed. The test in the Michigan Panel Study of Income Dynamics is unreliable, and was taken at the time the survey was administered (Mueser, forthcoming). Respondents in the NBER-TH sample were all in the military, and scored at or above the median (Taubman and Wales, 1974).

Because of its local nature, the Kalamazoo data does not remedy the need for large, representative samples with ability measures. That it adds significantly to available data reflects the meager base on which analyses in this area are conducted.

[7] One quarter of the respondents are from families in which more than two brothers were interviewed. Consequently, there are actually more than 346 unique pairs. I weighted the sample so that no individual would count as appearing in more than one pair.

[8] Featherman and Hauser (1975). I am grateful to Robert Hauser for making this information available to me.

[9] This speculation assumes that respondents' fathers who left Kalamazoo were disproportionately lower status. For support, at least for the early part of the century, see Thernstrom (1973). For a contrary view which emphasizes the greater success of out-migrants and in-migrants among the 1962 OCG respondents, see Blau and Duncan (1967) and

Duncan, Featherman and Duncan (1972).

[10]Eighty-two percent of the OCG II men 35 to 59 were between 35 and 54.

[11]U.S. Bureau of the Census (1975), Table 34.

[12]For discussion of this and other issues relating to differences in results across samples see McClleland (forthcoming).

[13]Intergenerational correlations are lower in the 1966 Detroit Area Survey than in the 1962 OCG I survey. See Duncan, Featherman, and Duncan (1972, p. 46).

[14]There is a disproportionate number of mangers, administrators and proprietors in the sample compared to the number in the total 1970 Kalamazoo male workforce aged 16 and over, and compared to the number in the 1970 Lansing, Michigan male workforce aged 35 to 54. See Olneck (1976a, p. 25).

[15]In a check in the Kalamazoo data, I found that respondents' reports of their brothers' educations had almost the same correlations with respondents' characteristics as did brothers' own reports of education. The degree of similarity between correlations involving Brother's Education in the Kalamazoo and OCG II samples would probably not be changed if OCG II had interviewed brothers.

[16]See Hildreth (1925); Corcoran, Jencks, Olneck (1976).

[17]See Jencks (1972); Hermalin (1969); Eaglesfield (forthcoming).

[18]See Behrman, Taubman, and Wales (1976); Corcoran, Jencks, and Olneck (1976). Restricting the Kalamazoo sample to pairs of brothers who differ in age by three or less years exaggerates rather than narrows discrepancies between correlations in the DZ portion of the NAS-NRC twin sample and the Kalamazoo brothers sample. Except for correlations involving ln earnings, the NAS-NRC DZ twin correlations tend to be appreciably lower than analogous correlations in the Kalamazoo sample.

[19]For regressions of son's outcomes on background measures see Olneck (1976a, forthcoming).

[20]Adding measures of maternal education, family composition, paternal nativity, father white-collar, and significant nonlinear and interaction terms raises the proportion of variance explained by measured background slightly, but never by more than 0.037 for any outcome. Consequently, I have used only three basic background variables in the present analyses.

[21]See Duncan, (1966).

[22]The predicted sibling correlations for Education, Initial Occupation, Current Occupation, Earnings, and Ln Earnings, taking into account sibling resemblance on test scores are 0.353, 0.264, 0.165, 0.090 and 0.082.

[23]Adding variables measuring family composition, race, and farm background never raises R^2 by greater than 0.022 in the OCG II data I analyzed.

[24]See Olneck (1976, Chapter 4) for these comparisons, and for the derivation of my corrections for measurement error.

[25]Jencks et al. (1972) report such comparisons for occupational status and income. See Inequality p. 201 and pp. 239-240. They erroneously refer to the formula for average sibling differences as 1.13 times the within-pair standard deviation. The formula which they actually give, i.e. $1.13\ [1\ rsib]^{1/2} S$, involves the within-pair standard deviation corrected for degrees of freedom. See Column 4 in Table 4 below.

[26]The standard deviation of predicted family means for earnings is $7634\ (.237)^{1/2} = 3716$. The standard deviation of earnings eliminating the effects of family background is 6668.

For similar comparisons for ln earnings in several data sets, see Corcoran, Jencks, and Olneck (1976).

[27]In a model predicting Occupation that takes into account the effects of Education and Initial Occupation, the correlation between the error terms for brothers is 0.284 for pairs 3 or less years apart in age, but only 0.001 for pairs more than 3 years in age.

[28]It is not due to the presence of outliers. I looked at cross-tabulations of brothers' Duncan scores categorized into 5 point intervals for the two groups. The number of pairs with very large differences in Duncan scores is similar for widely-spaced and closely-spaced brothers. In general, the spread of brothers' Duncan scores tends to be greater for all levels of respondents' scores for widely-spaced brothers than for closely-spaced brothers.

There is some suggestion that a similar conclusion might hold for earnings when brothers are very far apart in age. The correlation between earnings for brothers 5 or less years apart in age is 0.281, but it is only 0.108 for brothers more than 5 years apart. However, the difference between these correlations is not significant, and the correlation between absolute age difference and absolute earnings difference is only 0.054.

[29]For similar models which include initial and current occupational status see Olneck (1976a).

[30]Figure 2 is a variant of Figure B-7 in Jencks et al., (1972). I considered an alternative model in which orthogonal family background factors, one affecting all outcomes, one affecting all but the first outcome, one affecting all but the first two, and so on, are posited. In my data, the path to earnings from a factor common to test scores, education, and earnings is imaginary, so I abandoned the model. Nor did I estimate models in which measured background exercises direct effects, and unmeasured background factors are defined as orthogonal to measured background. I estimated the model shown below by hand calculation from observed correlations. Consequently, I cannot report standard errors for

the correlations among hypothetical variables.

[31]Olneck (1976a) reports, however, that inclusion of high school teachers' ratings of several personality characteristics such as industriousness, dependability, and executive ability, does not improve the prediction of sibling correlations on economic outcomes. See Olneck (1976a, Chapter 5).

[32]For a similar critique and an attempt to decompose the occupation-education relationship in Norway into familial and nonfamilial components, see Sweetser (1975).

[33]This conclusion should be generalized cautiously. It is not so strongly supported by Behrman, Taubman, and Wales (1976). Moreover, the effects of secondary education on initial occupation in the Michigan Panel Study of Income Dynamics, and in my data are smaller and less robust than the effects of higher education. See Olneck, 1976b. This is also true in the 1973 OCG II sample I analyzed.

[34]In Behrman, Taubman and Wales (1976) the within-pair education coefficient for DZ twins in the NAS-NRC sample is 92 percent as large as the uncontrolled coefficient. The cross-sibling correlation for education and occupation in that data is anomolously low compared to the analagous correlation in the Kalamazoo and OCG II data, so I tend to favor the Kalamazoo results. In the OCG II data, for 6865 respondents, 35 to 59, who reported their brother's education, controlling father's education, father's occupation, number of siblings, family composition, race, and farm background reduces the occupation-education coefficient by 15.0 percent. Using reports of brothers' education to calculate a within-pair occupation-education coefficient reduces the uncontrolled relationships by 23.2 percent. The importance of unmeasured compared to measured background factors for bias in the occupation-education relationship is less in the 1962 OCG I data than it is in the OCG II. See Olneck (1976b).

[35]See Griffin (1976) and Olneck (1976b). Olneck (1976b) assesses differential bias by level of schooling, and finds that the occupational effects of completing college are larger and more robust than the effects of completing high school.

[36]Antilog 0.1060 = 1.1118. A one standard deviation difference in test scores in the Kalamazoo data is associated with a 17.6 percent difference in earnings. A one standard deviation difference in test scores is associated with a 10 percent difference in 1971 earnings and a 5.7 percent difference in 1968 earnings among 1957 Wisconsin high school graduates (Hauser and Daymont, 1976), a 9.6 percent of difference in expected 1964 earnings among NORC Veterans respondents aged 25 to 34, and 17.5 percent among Veterans 30 to 34 (Jencks, forthcoming), and a 9.2 percent difference in 1972 earnings of Project Talent 11 year follow-up respondents (Crouse, forthcoming). These comparisons indicate that estimates of the effects of tested ability vary by <u>both</u> age of respondents and tests. This accounts, in part, for differences among researchers in estimates of the proportionate and absolute "ability" biases in the effects of education.

[37]Crouse (forthcoming a) offers little support for this interpretation however. The correlations between the separate components of the Project Talent Academic Composite and ln earnings do not differ significantly in the Talent 11 year follow-up.

[38]Controlling measured background in the NAS-NRC twin sample reduces the bivariate education-ln earnings relationship by 12 percent. The within-DZ pair coefficient is, however, only .059/.080 = 73.8 percent of the uncontrolled coefficient, and the within-MZ pair coefficient is only .027/.080 = 33.8 percent as large as the uncontrolled coefficient. The difference between MZ and DZ results suggests that either controlling genes is important, or that MZ twins share more common environments than do DZ twins. See Behrman, Taubman, and Wales (1976).

The 1973 OCG II data also suggest the importance of controlling unmeasured as well as measured background. Controlling measured socioeconomic background among 6855 respondents, aged 35 to 59, who reported their brother's education, reduces the relationship betweeen education and ln earnings by 19.7 percent. But using the correlations among respondent's education, respondents' ln earnings, and brother's education to calculate a within-pair coefficient reduces the relationship by 36.4 percent. The 1962 OCG I data do not, however, suggest dramatic differences between the education coefficients controlling measured background and unmeasured background. I have not yet investigated the possible sources of the discrepancy between the OCG I and OCG II results. See Olneck (1976b).

[39]The question may be raised as to whether it is more appropriate to estimate and compare proportionate or absolute biases across samples or within populations sampled longitudinally. If the uncontrolled effects of education differ between samples, the proportionate biases will differ even when absolute biases are the same. In longitudinal studies, if the effects of education rise faster than the effects of test scores or background, the proportionate bias will fall even though the absolute bias increases. It is probably best to report both absolute and proportionate biases. See Hauser and Daymont (1976), Griffin (1976), and Olneck (1976b).

Olneck (1976b) indicates that the observed effects of secondary schooling are more biased than the effects of higher education.

[40]Bishop estimated the correlation between reported and true values as 0.90, assuming that errors in separate reports of education are correlated 0.40 (Bishop, 1976; p. 5). I estimated the correlation between true and reported values of education in the Kalamazoo data as 0.964 (Olneck, 1976; pp. 172-178).

I calculated the error variance of schooling as $(2.73)^2 (1-0.964^2) = 0.5292$. Bishop gives the ratio of the observed to tne true coefficient as $b_t/\beta = 1/\alpha \; [1 - \frac{2V(ui)}{V(\Delta P)}]$ where

β = true coefficient

b_t = observed

α = correction for floor and ceiling effects producing a correlation between the errors in measurement and true values.

$V(u_u)$ = error variance in education

$V(\Delta P)$ = variance of sibling differences in education.

Adopting Bishop's values of = 0.95, I have bt/β = [1 - 2 (.5292)/6.720] ÷ .95 = .887.

[41]Assuming random errors and a reliability of 0.929, the error variance in schooling is $(2.73)^2(1-0.929) = 0.5292$. The ratio of error variance to the variance of sibling differences is 0.5292/6.7288 = .07865. If errors in test scores are random, assuming a reliability of 0.900 yields an error variance of $(15.32)^2(1-0.900) = 23.3292$. The ratio of error variance in test scores to the variance of sibling differences is 23.3292/249.5294 = 0.0935 (See Bishop, 1976).

[42]For the connection between status attainment research and American values see Blau and Duncan (1967,esp. pp. 432-441) and Jencks et al. (1972, Chapter 1).

[43]For an argument that genetic endowments explain substantial amounts of variance in earnings see Behrman, Taubman, and Wales (1976). For a critique of Behrman, Taubman, and Wales, see Goldberger (1976).

REFERENCES

Behrman, Jere; Paul Taubman, and Terence Wales, "Controlling for and Measuring the Effects of Genetics and Family Environment in Equations for Schooling and Labor Market Success," Paper prepared for presentation at the Mathematical Social Sciences Board Conference on Kinometrics, Williamstown, Massachusetts, May 1976.

Bishop, John, "Reporting Errors and the True Return to Schooling," Unpublished paper, Madison, 1976.

Blau, Peter, and Otis D. Duncan, The American Occupational Structure, Wiley, New York, 1967.

Bowles, Samuel, "Schooling and Inequality from Generation to Generation," The Journal of Political Economy, 80, May/June 1972.

Chamberlain, Gary, and Zvi Griliches, "Returns to Schooling of Brothers and Ability as an Unobserved Variance Component," Harvard Institute of Economic Research Discussion Paper 340, February, 1974.

Corcoran, Mary, Christopher Jencks, and Michael Olneck, "The Effects of Family Background on Earnings," American Economic Review, 66, May, 1976.

Crouse, James, "The Effects of Academic Ability," Draft, Chapter 10 in Christopher Jencks (ed.), Who Gets Ahead?, Basic, New York, forthcoming.

__________, "The Project Talent 11-14 Years Longitudinal Surveys," Draft, Appendix H in Jencks (ed.), Who Gets Ahead?.

Duncan, Otis D., "Ability and Achievement," Eugenics Quarterly, 15, March, 1968.

______________, "Path Analysis: Sociological Examples," The American Journal of Sociology, 72, July 1966.

______________, David Featherman, and Beverly Duncan, Socioeconomic Background and Achievement, Seminar Press, New York, 1972.

Eaglesfield, David, "The 1973 NORC Amalgam Survey," Draft, Appendix E in Jencks (ed.) Who Gets Ahead?

Fägerlind, Ingemar, Formal Education and Adult Earnings, Almquist and Wiksell, Stockholm, 1975.

Featherman, David L., and Robert M. Hauser, "Design for a Replicate Study of Social Mobility in the United States," in Kenneth C. Land and Seymour Spilerman (eds.), Social Indicator Models, Russell Sage, New York, 1975.

Goldberger, Arthur, "Twin Methods: A Skeptical View," Paper prepared for presentation at the Mathematical Social Sciences Board Conference on Kinometrics, Williamstown, Massachusetts, May 1976.

Gorseline, Donald E., The Effect of Schooling Upon Income, Graduate Council of Indiana University, 1932.

Griffin, Larry J., "Specification Biases in Estimates of Socioeconomic Returns to Schooling," Sociology of Education, 49, April 1976.

Griliches, Zvi and William M. Mason, "Education, Income, and Ability," Journal of Political Economy, 80, May/June 1972.

Hause, John C., "Earnings Profile: Ability and Schooling," Journal of Political Economy, 80, May/June 1972.

Hauser, Robert M. and Thomas N. Daymont, "Schooling, Ability and Earnings: Cross-Sectional Findings 8 to 14 Years After High School Graduation," Center for Demography and Ecology Working Paper 76-19, Madison, July 1976.

Hermalin, Albert, "The Homogeniety of Siblings on Education and Occupation," Unpublished doctoral dissertation, Princeton University, 1969.

Hildreth, Gertrude H., The Resemblance of Siblings in Intelligence and Achievement, Teachers College, New York, 1925.

Jencks, Christopher et al., Inequality: A Reassessment of the Effect of Family and Schooling in America, Basic, New York, 1972.

Jencks, Christopher, "The 1964 Veterans Survey," Draft, Appendix G in Jencks (ed.), Who Gets Ahead?.

McClleland, Kent, "Why Different Surveys Yield Different Results: The Case of Education and Earnings," Draft, Chapter 6 in Jencks (ed.), Who Gets Ahead?.

Mueser, Peter, 'The 1967-74 "Panel Study of Income Dynamics' Survey,' Draft, Appendix C in Jencks (ed.), Who Gets Ahead?.

Olneck, Michael R., "The Determinants of Educational Attainment and Adult Status Among Brothers: The Kalamazoo Study," Unpublished doctoral dissertation, Harvard Graduate School of Education, 1976.

______________, "The Effects of Education on Occupational Status and Earnings," Institute for Research on Poverty Discussion Paper 358-76, Madison, August 1976. Draft, Chapter 12 in Jencks (ed.), Who Gets Ahead?.

____________, "The Kalamazoo Brothers Sample," Draft, Appendix I in Jencks (ed.), Who Gets Ahead?.

Sewell, William H. and Robert M. Hauser, Education, Occupation, and Earnings, Academic Press, New York, 1975.

Sweetser, Dorian Apple, "Education and Privilege: Analyses of Sibling Occupational Mobility," Acta Sociologica 18, 1975.

Taubman, Paul and Terence Wales, Higher Education and Earnings: College as an Investment and a Screening Device McGraw-Hill, New York, 1974.

Thernstrom, Stephen, The Other Bostonians, Harvard University Press, Cambridge, 1973.

U.S. Bureau of the Census, "Money Income in 1973 of Families and Persons in the United States," Population Reports, Series P-60, No. 97, U.S. Government Printing Office, Washington, D.C., 1975.

Welch, Finis, "Relationships Between Income and Schooling," in F.N. Kerlinger and J. B. Carroll (eds.), Review of Research in Education, Peacock, Itasca, Illinois, 1974.

FAMILY BACKGROUND AND ACHIEVEMENT: A COMMENT

Robert T. Michael
Stanford University and NBER

Olneck's paper makes an important contribution to the rapidly growing literature on the relationship of family background to subsequent achievement. With entrepreneurial effort matched by few but the editor of this volume, Olneck identified, collected and analyzed a quite useful source of data on male siblings who attended grade school in Kalamazoo, Michigan over a period of two decades (1928-1950). He is to be commended (and assuming he shares his data with others, thanked) for his efforts in obtaining the data set. Olneck's analysis produced several findings which are informing and which should prove useful in subsequent study.

Olneck's most important finding, which provides independent support for results from other investigations, is that brothers are more alike in earnings, occupation and schooling attainment than can be explained by measured background characteristics or by measured similarity in ability test scores and other frequently used explanatory variables. Several studies have shown us that a sizable portion of the variance in, say, earnings (or log earnings) cannot be attributed to measured variables such as job experience, schooling, and ability; yet other studies of longitudinal data on earnings have indicated that there is a positive covariation in the unexplained earnings residual for an individual from year to year. This evidence suggests that there are persistent and perhaps systematic forces affecting earnings which have not yet been identified.

Olneck's evidence provides a valuable clue that family-related factors are in part responsible for these forces. These factors work through channels other than those reflected in standard recursive models of earnings. Olneck does not provide any hints about what these channels are, be they motivation, attitude or skills, social status, wealth, access to capital, connections, or whatever.

Clues are of value for their potential, however, and in my judgment the latent variables approach with which Olneck exploits his clues about the importance of family background is not as useful as a direct approach investigating what measurable background characteristics appear to explain the similarities between brothers. Olneck states that "the data do not enable us to determine what it is that brothers share which accounts for" the family background effects, but he bears much responsibility for what information this data set does and does not contain. Information about whether the family effects result from characteristics about which the parents had no control (e.g., their race, their age at some point in time), had some control at one time (e.g., selective mating variables, parental schooling levels, number of offspring), or had control during the brothers' childhoods (e.g., resources expended on the brothers, mother's labor market behavior and parents' time allocation in general) might have proven very useful. Even negative conclusions would have been of interest by suggesting restrictions on the channels through which family effects might have operated.

Turning to specifics, sample attrition was severe if judged in terms of Olneck's potential sample — less than 60 percent of the men were traced and of these only 43 percent are included in the sample analyzed in the bulk of the paper. It seems likely that the differences between brothers in characteristics such as earnings would be greater for pairs not located. Apparently, much of the searching for these brothers was limited to a local geographical area and brothers who move away as well as those who, for example, undertake more unusual occupations may be expected to be both more difficult to locate and more disparate in schooling attainment and earnings. So it seems to me the data are probably censored with respect to the dependent variables which imposes bias on the estimated coefficients. Despite Olneck's careful and useful comparison of his Kalamazoo sample with OCG II and other data, I think a nonrandom sampling of only 25 percent of a sample pool is quite troublesome.

Several times in his paper Olneck notes the dispersion in earnings between brothers, after standardizing for test scores and schooling, and emphasizes the "capriciousness and randomness which characterize the economic game." But age differences between the brothers are not eliminated. Given the positive slope of the typical age-earnings profile for men between ages 35 and 50, considerable dispersion would be expected even if the two brothers were on an identical deterministic age-earnings path.[1] Moreover, the purely "random", year-to-year fluctuations in earnings tend to cancel out over time and a several-year-average earnings would show less fluctuation, greater relative association with the personal characteristics studied, and be of more relevance for many purposes, including the measurement of income inequality which Olneck discusses.

Another theme evident throughout Olneck's paper is that unmeasured family background factors may impart biases to estimated coefficients reflecting, say, the influence of ability or schooling on earnings. However, biases are not intrinsic; they exist in the context of some particular model or in estimating some well-specified relationship. If one is interested in a statistic which estimates the parameter reflecting the effect of y on x holding z constant, $r_{xy.z}$, then the statistic r_{xy} is biased if $\rho_{yz} \neq 0$ or $\rho_{xz} \neq 0$. One may wish to estimate the effect without holding z constant, in which case $r_{xy.z}$ is "biased." The point is an obvious one, but its application to the statistics with which Olneck is working is less than obvious.

For example, we are told that in the sample as a whole a 10-point increase in test score is related to 1.0 additional years of schooling whereas within a family a 10-point difference is related to 0.6 additional years of schooling. So if society were able to hand a child 10 additional test score points (and whatever substance it reflects), what do we expect to be the impact on the child's schooling? Well, compared to what? Compared to what we might have expected for the child who was drawn randomly from the Kalamazoo sample, 1.0 years. Compared to what we would have expected for the child in this child's family, 0.6 years assuming his parents are permitted to reallocate resources as they in fact do and assuming his brother is a good proxy for himself. In many instances pertaining to both schooling and earnings, the conditions for which one wants an estimated effect are not obvious, so discussions about biases require precision.

Furthermore, even if it is clear that one wants an estimate of the parameter $\rho_{xy.z}$, within-family estimation is appropriate only if the correlation between y and z is weaker within families than among families.[2] If z is, say, motivation, the correlation between y (test score) and z may be higher within the family in which case a within-family estimate of $\rho_{xy.z}$ is more biased than an across-family estimate. Olneck's estimates of these relationships between test score and earnings are interesting and somewhat puzzling. They suggest that within families z has either opposite-signed correlations with x and y or less strong same-signed correlations. (Olneck suggests one example of the latter.)[3]

The likelihood that parents attempt to offset differences among their children or to reinforce those differences is related to my principal concern with Olneck's paper, the notion of family background. Family background is defined by Olneck as, "all those factors, both measured and unmeasured, which produce resemblance on outcomes among siblings." But "all those factors" are so closely related to the selection of the sample that the definition yields an arbitrary delineation of what is and is not family background -- had his sample contained families in other cities or regions, of other religions, incomes or family structures, he would have included cross-sectional differences among families in school quality, in climate (political, social and meteorological), in relative prices of market goods and services, et cetera as family background; whereas, had he limited his sample to a single five-year interval instead of a twenty-two year span, he would have removed whatever cohort effects exist in growing up in the late 1920's instead of the post-WW II 1940's. To define family background as whatever produces resemblance, when resemblance is itself a relative notion, offers no guidance about just what is and what is not included in the concept.

More importantly, to measure the effect of family background as the resemblance between brothers assumes parents are passive conduits through which "background characteristics" affect their children. But surely parents exercise some volition about their influence on (or at least the resources devoted to) each of their children. Confronted with two children with different personalities, energy, motivation, intelligence, health, et cetera, parents presumably make decisions about the level and composition of resources (time and money) they will devote to each child. Only if it is the parents' intention to make their children as similar as possible (in terms of tested intelligence, schooling level, occupation and earnings) does Olneck's analysis reveal the influence of parents' characteristics on their children.

If parents respond to differences in children's attributes by promoting the comparative advantages of each child (perhaps even devoting different levels of total resources among children on the basis of some sort of absolute advantage), then similarity between brothers is not evidence of the importance of family characteristics nor is dissimilarity evidence of its lack of importance. Indeed, if parents respond so as to reinforce the potentials of gifted children (e.g., mentally, musically, physically gifted), then it may be in families in which there is the widest difference in achievement between siblings that there is in fact the greatest influence of family background.[4]

In discussions of family background in the literature, many studies investigate the effects of explicit measures of parental or family characteristics. Olneck points out that several of these explicit measures together do not account for the similarities between brothers in his data. This fact is quite interesting as are estimates of the gross magnitudes of the resemblances between brothers and the net magnitudes conditional upon differences in intervening variables. However, it is not informative to label these resemblances "family background" effects, nor to ignore the difference between the influence on children of what parents are and what parents do. It may seem unfair to ask of Olneck insights about the resemblances between brothers when he has done a considerable service in providing estimates of their magnitudes. The point applies to most of the "kinometric" research; the estimated numbers are interesting, but the interpretations are far from straightforward.

In his dialogue, The Republic, Plato asks, "and how shall we manage the period between birth and education, which seems to require the greatest care? Tell us how these things will be." While I think none of us would endorse his substantive answer, we might sense wisdom in his initial response: "Yes, my simple friend, but the answer is the reverse of easy; many more doubts arise about this than about our previous conclusions."

FOOTNOTES

[1] The magnitude of the slope of the profile is greater at higher levels of schooling in this age interval. Judging by figures for all males with income from the 1966 C.P.S., in the five years between age 35 and age 40 income (not earnings) increases by about 9 percent (about $800) for men with the schooling levels average in the Kalamazoo sample and by about 14 percent ($1600) for men with sixteen or more years of schooling. (Current Population Reports, P-60, No. 56; August 14, 1968). Olneck is aware of this issue and discusses it in his section on "spacing." He indicates that absolute differences in earnings between brothers are unrelated to age differences, and his Table 5 shows that the within-pair standard deviation in earnings is higher for brothers who differ in age by three years or less compared to those who differ in age by more than three years, but the sibling correlations are higher for the former group than for the latter group. I find this result puzzling; it doesn't appear to be related to a bigger difference in schooling levels for brothers who are closer in age.

[2] This argument is extensively developed in a related context in G.S. Becker and N. Tomes, "Child Endowments and the Quantity and Quality of Children," J.P.E., 84, No. 4, part 2 (August 1976) p. S157.

[3] If parents attempt to offset ability differences in order to achieve more similar schooling levels among their children, this could explain a negative relationship between y and z, but it seems an unlikely explanation.

[4] If children were virtually identical when they entered a family, I know of no theoretical reason why parents would behave so as to create differences. But children within a family are indisputably different in many respects, and there are theoretical reasons to expect parents to respond to these differences. Becker and Tomes (1976) explore these incentives analytically and suggest it is likely that parents tend to invest human capital in their children in a manner which reinforces differences, but to transfer (any) nonhuman capital to children in a way that compensates for differences in skills.

CHAPTER 6

GENES AND SOCIAL STRATIFICATION
A Methodological Exploration with Illustrative Data

Christopher Jencks*
Department of Sociology
Harvard University
and
Marsha Brown*
University of Washington

The first section of this paper discusses the recent "heredity/environment" debate and argues that it is based on a false dichotomy. That dichotomy assumes that genes act independently of the environment. In fact, an organism's genes influence its development partly by influencing its environment. Thus even if genetic variation explains all the observed variance in a given trait, one cannot conclude that variations in the environment do not affect the trait. One can only conclude that genetic factors explain all variation in the relevant environmental factors. Usually genes explain only a fraction of the observed variation in a given trait, and one does not know whether this is because the genes in question affect the individual's environment. In this case, the explanatory power of genotype sets no limit whatever on the effects of the environment. High correlations between phenotype and genotype may, however, provide useful clues about the mechanisms by which a given trait is determined. Students of social stratification can therefore learn something from estimates of the correlation between genotype and such traits as test performance, educational attainment, occupational status, and earnings.

Section II describes the sample we will use to estimate some of these correlations. The sample covers 1203 pairs of twins and siblings surveyed by Project Talent. The respondents took a wide range of cognitive tests in 1960, when they were enrolled in grades 9 through 12. They were followed up in 1965-68, five years after their class had graduated from high school. Since many were still enrolled in college or graduate school at this time, and since many others had temporary jobs (including jobs in the military), their current occupational statuses and earnings tell little about their eventual success. The five-year followup does, however, provide quite reliable estimates of eventual educational attainment. It also provides a measure of each respondent's career plans.

Section III presents formulas for using twin data to estimate the

*We are indebted to James Crouse, David Featherman, William Kruskal, John Loehlin, Michael Olneck, and Paul Taubman for comments on earlier drafts. The Carnegie Corporation of New York supported our work on these issues from 1971 to 1973. The views we express here are, however, entirely our own.

percentage of variance in human traits explained by variations in genotype. It also discusses the substantive assumptions behind these formulas and assesses the empirical evidence for and against them.

Section IV uses the data presented in Section II and the formulas in Section III to estimate the "heritability" of test performance, educational attainment, and career plans. Using conventional assumptions, these heritabilities are all significantly greater than zero, but their standard errors are also quite large, and the results of career plans are internally contradictory.

Section V investigates the extent to which the genes that influence educational attainment operate by influencing test scores. The data imply that genes have substantial effects on education that are independent of test performance. This suggests that investigators who restrict themselves to estimating the impact of IQ genotype on adult success, as Jencks et al. (1972) did, may seriously underestimate genes' overall impact on success.

The conclusion reviews some of the methodological problems encountered in earlier sections and summarizes our substantive findings.

I. Conceptual Issues

We will use three terms from population genetics: phenotype, genotype, and heritability. A simple example will illustrate their meaning. Suppose a farmer alternates two genetically pure strains of corn in the same field. Suppose that after many years of alternating strains in this way he finds that Strain A has yielded an average of 1120 bushels per year, with a standard deviation of 160 bushels, while Strain B has yielded an average of 880 bushels per year, again with a standard deviation of 160 bushels. Under these circumstances we can define our terms as follows:

Phenotype is the observed yield in any specific year. The phenotypic mean for all years is (1120 + 880)/2 = 1000 bushels. The between-strains phenotypic variance is $[(1120 - 1000)^2 + (880 - 1000)^2]/2 = 120^2 =$ 14,400. The within-strains phenotypic variance is $160^2 = 25{,}600$. The total phenotypic variance is therefore 14,400 + 25,600 = 40,000.

Genotype is the average yield of a given stain when it encounters the range of environments provided by this particular field over the years in question. The genotypic values are 1120 bushels for Strain A and 880 bushels for Strain B. The genotypic mean for both strains together is by definition the same as the phenotypic mean, i.e. 1000 bushels. The genotypic variance is the variance of the genotypic values. This is simply the between-strains phenotypic variance, i.e. 14,400.

Broad Heritability (h^2) is the percentage of the phenotypic[1] variance that can be explained (in the statistical sense) by genotype. This is the same as the ratio of the genotypic variance to the phenotypic variance. In this case h^2 = 14,000/40,000 = 0.36.

Many unwary investigators treat $1 - h^2$ as an estimate of the fraction of the phenotypic variance attributable to environmental variation. This is an extremely misleading practice. It is true that in this example since each strain is genetically homogeneous at least 25,600/40,000 = 64 percent of the phenotypic variance occurs within strains and must therefore be due to non-genetic factors. But the remaining 36 percent, while clearly traceable to genes, may also be traceable to variations in the environment. Suppose, for example, that yield depends entirely on the percentage of the crop attacked by pests, and that Strain A has a higher yield than Strain B solely because it is less attractive to pests. Pests are clearly part of the environment as most people conceive it . Equalizing the environment of the two strains therefore implies equalizing the number of pests attacking each strain. But if this were done the two strains would have equal yields. Conventional usage, then, would imply that environmental factors explain 100 percent of the variance in yields. This is not to deny that genes explain 36 percent of the variance. The point is rather that genes explain 36 percent of the phenotypic variance because they account for 36 percent of the variation in a plant's effective environment.

Now let us turn to a human example. Human beings normally have 23 pairs of chromosomes. Each pair of chromosomes contains thousands of genetic "loci". An individual has two genes at each locus: one from his father, one from his mother. In the absence of mutations, the gene received from the father is always identical to one of the father's two genes at the same locus, and the gene received from the mother is identical to one of the mother's genes. Thus, if the father has genes A and B at a given locus, while the mother has genes C and D at the same locus, the child is equally likely to be an AC, AD, BC or BD.

Traits such as social position are affected by many different genes, only a handful of which have been identified. This means we cannot assume two individuals have the same genotype for a given trait unless we know that all their genes are alike. The odds against two individuals ending up with the same genes by chance alone are overwhelming. The odds can be sharply improved through intense inbreeding, but incest taboos dicourage this, and even if they did not, intense inbreeding seems to reduce fitness as the population approaches homozygosity.

Fortunately for genetic researchers, we have another potential source of genetically identical individuals. Roughly one American birth in 80 results in twins. Roughly two-thirds of these twins are dizygotic (DZ), meaning they come from two different eggs, fertilized by two different sperm. Such twins are no more alike genetically than ordinary siblings. Roughly a third of American twins are monozygotic (MZ) , meaning they come from an egg that has divided after fertilization. Barring mutations, which are extremely rare, these twins have precisely the same genes.

Now suppose we collect data on a large sample of identical twins who have been separated at birth. Suppose we could somehow satisfy ourselves that these twins' prenatal environments were no more alike than those of random individuals,[2] and that adoption agencies had assigned the members of each pair to new homes on a random basis. Under these circumstances any correlation between twins' traits would be attributable to their common genes. This correlation would be the same as the ratio of the genotypic variance to the total variance for twins. Thus if the correlation between separated twins' earnings were 0.25 and if the variance of earnings were $(\$10,000)^2$, the variance of twins' genotypes for earnings would be $(0.25)(10,000)^2 = 5,000^2$. If the twins were genetically representative of the larger population, and if the variance of earnings were $\$10,000^2$ in the larger population, we could infer a heritability of 0.25 in this population. If the variance of earnings were, say, $(\$12,000)^2$ in the larger population, the implied heritability would be $5000^2/12,000^2 = 0.17$.[3]

The fact that genes explained 25 percent of the variance in earnings would not necessarily mean that "environment" explained only 75 percent. If our twins were representative of the American population, for example, they would be roughly half male and half female. Since genetically identical twins are necessarily of the same sex, and since sex explains something like 25 percent of the variance in adult earnings, identical twins' earnings would necessarily correlate at least 0.25. But while a quarter of the variance in earnings may depend on whether an individual has an XX or an XY chromosome, this variance may also be explicable in environmental terms. Sex affects the kinds of opportunities employers provide, the expectations spouses have for one another, and a host of other environmental factors. If these environmental differences were eliminated, sex differences in earnings would fall. Thus while our hypothetical twin study would prove that genes explained 25 percent of the variance in earnings, it would not prove that environment explained only 75 percent of the variance. Environment might well explain 100 percent of the variance.

In an earlier effort to deal with this problem Jencks et al., (1972) redefined "environment" so that it included only those aspects of the environment that did not depend on genotype. Jensen has adopted this same expedient. Instead of saying that equalizing the environment means equalizing the number of pests attacking Strains A and B, for example, one says that equalizing the environment implies equalizing the numbers of pests that "could" attack each strain. This requirement is presumably met if the two strains have been alternated in the same field over a sufficiently long period of time. Any differences in the mean yield of the two strains are then defined as "genetic" not "environmental." Applying this logic to humans forces us to use words in rather bizarre ways, however. Consider the case of sex differences. Males and females are born into the same families. Differences in their eventual earnings are therefore by definition potentially traceable to the fact that they have different genes. If they have different environments, that is because they have different genes. If we define "environment" so as to exclude those aspects of environment that depend on genes, sex differences can only contribute to the genetic variance, not the environmental variance. The same holds for race differences. Some investigators try to get around this problem by looking only at white males, or by stratifying their sample on sex and race. But this solution is only reasonable if the other genes that affect adult success exert their influence in ways that are completely different from the way sex and race genes affect success, i.e., if they do not affect the environment. Since we have no idea what these other genes are, or how they influence economic success, it is hard to justify either the assumption that they work in the same way as sex and race or the assumption that they work differently.

Behind this uncertainty about whether genes affect phenotypes by affecting an organism's environment lies a deeper uncertainty about what we really mean by environment. Almost everyone would agree that if an individual's genes influence the way other people treat him, the results are due to "environment" as well as "genes". But suppose genes influence an individual's own choices. Suppose, for example, that genes affect the ease with which different individuals learn to use new words. Suppose that individuals who master new terms easily find schooling pleasanter than those who master new terms with difficulty, and that those who enjoy school choose to stay in school longer than those who hate it. We might say that if two individuals attend the same school but react differently because they have different genes, the two individuals have the same environment and that phenotypic differences between them are due to the "direct" effects of genotype. But if we were to compare the daily lives of individuals who mastered new terms easily with the daily lives of those who mastered new terms with difficulty we would almost certainly find that they had different friends, enrolled in different courses, engaged in different leisure activities, and so forth. Even when they were nominally doing the same thing, e.g. attending a 10th grade English class, they would probably be directing their attention to different matters within the classroom. This argument suggests that when one looks sufficiently closely, individuals with different genotypes almost never encounter the same environment.

Conventional definitions of "environment" are thus fundamentally different from our definition of genotype. We can assert that two organisms have the same genes, and hence the same genotype, even if we do not know their phenotypic values on the trait that interests us. We made this assertion both for inbred strains of corn and for identical twins. We could also make it for other individuals if we had physically compared each gene at each locus, although our chances of finding individuals with identical genes at every locus would be infinitesimally small. We can never assert that two organisms have encountered the same environment, or even that the effects of their environments have been equal, unless we actually know their phenotypic values for a given trait. If two individuals have the same genes and also have the same phenotype, we assume that their environments "must" have had the same effect on this outcome. If two genetically identical individuals have different phenotypes, we assume that their environments "must" have differed. This usage effectively converts "environment" into a residual category that accounts for discrepancies between genotype and phenotype. There is nothing wrong with such residual categories. But there is no justification for assigning them misleading descriptive labels like "environment". In any other context social scientists would use a deliberately ambiguous label, such as "unmeasured factors" or simply "error" to describe such a residual. That seems preferable here, too. If a more precise label is needed, perhaps "non-genetic factors" would be appropriate.

Nonetheless, we cannot ignore environment entirely. Our definition of genotype assumed, for example, that our two strains of corn had been raised in the same fields in alternative years. Likewise, our estimates of h^2 for twins' earnings required data on twins who had been adopted by random parents. The estimates assumed, in other words, that in the absence of genetic differences, the expected phenotypic values must be identical. We will describe this situation as one in which environments are "initially similar". If this condition is not met, the observed between-strains variance for corn will no longer measure the effects of genotype on phenotype. Rather, it will measure the combined effects of genotype and initial differences in environment. If our hypothetical farmer had been able to anticipate the likely incidence of pests in each upcoming year, for example, and if he had responded by planting the strain that was less attractive to pests in years when the expected incidence was unusually high, the between-strains variance would have been less than 36 percent of the total variance. Indeed, it might have been zero. Conversely, if the farmer had planted the less attractive strain in years with a low predicted incidence of pests, the between-strains variance might have been more than 36 percent of the total. The same rule holds for our twins. If rich families adopted most of the boys, the correlation between sex and earnings would exceed 0.50. If rich families adopted girls, the correlation would be less than 0.50.

Past experience suggests that many readers will find this line af argument quite confusing. If heritability estimates give the percentage of the phenotypic variance explained by genotypes in situations where the genotypes have been initially allocated to random environments, many readers assume that this is equivalent to saying that heritability estimates only hold where genotypes are uncorrelated with environment. This is not the case. Our definition requires only that genotypes start life in random environments. It does not require that they remain in random environments. Insofar as the organism either chooses or creates its own environment, genotypes can and will end up correlated with environments.

We can formalize this distinction by dividing the sources of environmental variation into those that are "endogenous" (i.e. explained by genotype) and those that are "exogenous" (i.e. unaffected by genotype). Real features of the environment seldom fall neatly in one category or the other. If one is studying the determinants of a child's verbal ability, for example, the quality of the father's interactions with his children is likely to emerge as an important determinant of the child's verbal development. The quality of these interactions is partly exogenous, since it depends on such factors as the father's education and occupation, as well as the father's own genotype. But the quality of the father's interactions with his children is also partly endogenous, since it depends partly on how the child responds to the father. If the child is articulate and asks a lot of questions, the father will behave differently than if the child in inarticulate and withdrawn. This holds for other behavioral measures as well. The distinction between endogenous and exogenous features of the environment is thus heuristic rather than practical.

It may help to state this argument is algebraic terms. Let us scale all variables to means of zero. Let us then denote phenotype as P, genotype as G, exogenous features of the environment as X, and overall environment as E. Our model assumes two equations:

(1) $P = aG + bE + eW$

and

(2) $E = cG + dX + fV$

where a,b,c,d,e, and f are scalars (regression coefficients) and V and W are error terms. W embodies the non-additive effects of G and E on P, while V embodies the non-additive effects of G and X on E. Some may therefore prefer to think of W and V as interaction terms rather than as error terms. Figure 1 portrays this model visually.

We are interested in measuring the total effect of G on P when G and X are uncorrelated. This is the same as the total effect with X controlled, i.e. a + b c. As we shall see, one can estimate this effect by taking advantage of the fact that genetic theory allows us to estimate the correlation between certain individuals' genotypes with considerable confidence. We know, for example, that the correlation between identical twins' genotypes is 1.00. The same is true for members of sufficiently inbred strains. Likewise, we know that the correlation between the genotypes of ordinary siblings is typically on the order of 0.50, though this value may be inflated by assortative mating or depressed by non-additivity (i.e. dominance and epistasis). Section III presents methods for using these facts to estimate the effect of G on P.

In principle, we are also interested in estimating the effects of E on P, but in practice we cannot hope to do this. Neither psychological nor sociological theory provides any strong basis for believing we can actually measure E. Nor do psychologists or sociologists have any a priori basis for predicting the correlation between values of E for particular sets of relatives. E is thus a completely free endogenous parameter. There are no known constraints on its behavior. Under these circumstances one cannot hope to estimate its correlation with anything except by making completely arbitrary assumptions. As a result, we can only estimate a "reduced-form" model, in which E does not appear, i.e.

(3) $P = aG + b(cG + dX \pm fV) + eW = (a + bc)G + bdX + bfV + eW.$

Setting $a+bc = h$, $bd = i$, and $bfV + eW = jU$, we have

(4) $P = hG + iX + jU.$

Figure 2 displays this model visually. We use roman type to denote unstandardized values and italics to denote standardized values, i.e. values that have been rescaled to have standard devitations of 1.00 as well as means of zero. We have scaled G so that h = 1.00 by definition. It follows that $\underline{h} = (1.00)\ (s_G/s_P)$, hence that $\underline{h}^2 = s_G^2/s_P^2$ = "heritability."

How much variance is explained by "environment" (E)? The answer depends on precisely what one means by the term "explained." One possibility is to ask how much of the observed variance in P we could explain statistically if we knew E, i.e. how large R^2_{PE} is. If we standardize all variables and use italics to denote standardized coefficients,

(5) $R^2_{PE} = (\underline{b} + \underline{ac})^2 = \underline{b}^2 + 2\underline{abc} + \underline{a}^2\underline{c}^2.$

If we assume $R^2_{PE} = 1 - \underline{h}^2$, how wrong will we be? First, note that:

(6) $\underline{h}^2 = (\underline{a} + \underline{bc})^2 = \underline{a}^2 + \underline{b}^2\underline{c}^2 + 2\underline{abc}$

Furthermore,

(7) $1 = \underline{a}^2 + \underline{b}^2 + 2\underline{abc} + \underline{e}^2.$

Subtracting equation 6 from equation 7 we get:

(8) $1 - \underline{h}^2 = \underline{b}^2 - \underline{b}^2\underline{c}^2 + \underline{e}^2$

The bias that arises from assuming $R^2_{PE} = 1 - \underline{h}^2$ is then the difference between equation 5 and equation 8, i.e.

(9) $R^2_{PE} - (1 - \underline{h}^2) = 2\underline{abc} + \underline{a}^2\underline{c}^2 + \underline{b}^2\underline{c}^2 + \underline{e}^2$

Thus if the correlation between genotype and environment ($\underline{c}$) is positive, and if the effects of G and X are nearly additive (i.e. $\underline{e}^2$ is small), $1 - \underline{h}^2$ will underestimate R^2_{PE}.

Alternatively, one might want to know what percentage of the observed variance would disappear if environments were equalized. Equalizing environments makes $s_E = s_W = 0$. The phenotypic variance will therefore be $a^2 s^2_G$, or in standard form, $\underline{a}^2$. We can call this the "non-environmental" variance. The "environmental" variance is then $1 - \underline{a}^2$. If we assume $1 - \underline{a}^2 = 1 - \underline{h}^2$, the bias is equal to:

(10) $(1 - \underline{a}^2) - (1 - \underline{h}^2) = \underline{h}^2 - \underline{a}^2 = \underline{b}^2\underline{c}^2 + 2\underline{abc}.$

Using this definition, then $1 - \underline{h}^2$ overestimates the "environmental" variance if the correlation between G and E (i.e. $\underline{c}$) is positive, regardless of the size of e^2. But if society allocates environments so as to compensate for genetic disadvantages, $\underline{c}$ will be negative. In that case $1 - \underline{h}^2$ may over-estimate the effects of environmental inequality, though equations 9 and 10 show that this is not a foregone conclusion.

One final observation may be helpful. Debate over the heritability of certain human traits, notably "intelligence", has traditionally been highly political. Both liberals and conservatives seem to assume that high heritabilities help justify the status quo. This judgment is predicated on the notion that genetic variation is more difficult to eliminate than non-genetic variation. Thus, if most of the phenotypic variance in a trait is traceable to genetic variation, one can argue that egalitarian reform is doomed to failure.

This line of reasoning is wrong. Until one specifies the mechanisms by which genes affect human traits, high heritability has no political implications whatever. If heritability is high because of racism and sexism, for example, it can clearly be reduced by changing the norms that support discrimination against blacks and females. High heritabilities of this sort certainly do not justify the status quo or prove that economic inequality is inevitable. If the heritability of earnings is high because employers pay more for workers who are tall, strong, or free from inherited diseases, and if these traits are related to productivity, one can make a somewhat stronger case for the notion that high heritabilities imply that egalitarian reforms will fail. But even here the case is ultimately unconvincing. There is no 11th Commandment requiring that living standards be proportionate to economic productivity. Most societies, for example, encourage healthy adults to share their incomes with the young, the old, and the sick. Many make such sharing compulsory. Relatively few industrial societies requires much sharing among healthy adults whose productivity differs, but that is a matter of collective moral judgment, not a technical necessity.

These considerations suggest that if we want to draw any ethical or political conclusions from heritability data, we need to develop causal models of how genes influence a given outcome. While we will not be able to make much progress in this direction in this paper, Section V suggests some methods by which one can at least make a beginning.

II. Project Talent Data

The data analyzed in this paper come from Project Talent, a longitudinal survey of a stratified probability sample of American students enrolled in grades 9 through 12 in 1960. The Talent sample has been described extensively elsewhere,[4] and we will not repeat the details here. The initial sample of 400,000 students should have been quite representative of the target population. All members of the sample took a lengthy battery of tests and filled out a long questionnaire in 1960. Project Talent tried to resurvey all respondents by mail in 1965-68, five years after they would normally have finished high school. About 125,000 individuals responded to this followup. When we are interested in the determinants of educational attainment and career plans at 23, these are the only respondents with the required data. We will look at two subsamples: one composed of twins and one composed of ordinary siblings.

A. The Twin Sample

Schoenfeldt (1967, 1968) first assembled and analyzed the Talent twin sample. He found that 5.5 percent of the 1960 sample said they were twins, triplets, or quadruplets. Since only about two percent of all live births are twins (Loehlin and Nichols, 1976: 7), and a negligible percentage are triplets or quadruplets, he inferred that many of the responses were spurious. As one might expect, the percentages were more reasonable for older students and for girls. But even among 12th grade girls, 2.7 percent claimed to have a twin.[5] In order to identify genuine pairs, Schoenfeldt sorted all respondents who said they were twins by school and then scanned each school list visually for pairs with the same last name, the same address, and the same birthday. This procedure eliminated four-fifths of the self-reported twins, leaving 1,959 pairs. This is only 1.0 percent of the full sample, or half as many pairs as birth records might lead one to expect. There are several possible reasons for this. First, Schoenfeldt scanned school records visually and may have over-looked as many as a sixth of all eligible pairs.[6] Second, some respondents must have had a twin in another school, a twin who had quit school, a twin not enrolled in grades 9 to 12, or a twin who had died. (Mortality is unusually high among twins, especially males.) Still, the number of pairs is disturbingly low.

Another problem is that only 22.9 percent of Schoenfeldt's pairs were of opposite sex. Yet birth records indicate that 33 percent of all twins are of opposite sex.[7] Since roughly half of all DZ twins are of opposite sex, while no MZ twins are, Schoenfeldt's results imply that about 45.8 percent of the Talent twins were dizogotic, compared to 67 percent of all twins born in the United States. Perhaps dizygotic pairs more often include at least one twin who dies young, leaves school young, or is in a different grade or a different school. Or perhaps Schoenfeldt overlooked more mixed-sex than same-sex pairs.

Schoenfeldt sent all same-sex pairs enrolled in the same grade a questionnaire designed to determine their zygosity. He did not send questionnaires to pairs whose members were enrolled in different grades. Nor did he record the number of pairs eliminated for this reason. He got responses from both members of 616 pairs, diagnosing 423 pairs as monozygotic (MZ) and 193 pairs as dizygotic (DZ). This imbalance between MZ and DZ pairs is as expected, given the initial ratio of mixed-sex to same-sex pairs. Schoenfeldt (1967) reports that when he applied the same decision rules to another twin sample for which blood samples had been used to determine zygosity, he misdaignosed 7 out of 82 MZ pairs and 8 out of 42 DZ pairs. Thus 8/(82-7+8) = 9.6 percent of all pairs diagnosed as MZ were in fact DZ, while 17.1 percent of all pairs diagnosed as DZ were MZ.

In addition to the 616 same-sex pairs, 448 mixed-sex pairs were available, creating a total of the 1064 pairs with known zygosity. Five-year followup questionnaires were available for both members of 366 pairs, but 30 of these pairs had incomplete test data or incomplete educational data. Table 1 shows the distributions of the final sample by sex and zygosity.

B. The Sibling Sample

The sibling sample was drawn from 98 of Talent's public high schools and their associated junior highs.[8] It included all students who

Table 1

Sources of Attrition in Final Talent Twin Sample

	MZ Males	MZ Females	DZ Males	DZ Females	DZ Mixed Sex	Total
1. Estimated pairs per 400,000 live births	677	656	677	656	1333	4000
2. Estimated pairs in Schoenfeldt's sample	540	523	228	220	448	1959
3. Pairs of known zygosity	184	239	66	127	448	1064
4. Pairs with composite test scores and education data	76	90	26	51	93	336

Notes on Table 1

Line 1 assumes that one U.S. birth in 100 produces twins, that a third of these twins are identical, that 50.8 of all same-sex twins are male, and that 49.2 percent are female. See Loehlin and Nichols (1976), pp. 7-8.

Line 2. The N's in columns 5 and 6 are from Schoenfeldt (1968). To estimate columns 1-4, we assumed that the N for DZ same-sex pairs was equal to that for DZ opposite-sex pairs, and that the sex ratio was 50.8:49.2. We then assumed that all remaining pairs were MZ. These assumptions may not hold for twins in grades 9-12 of the same school.

Line 3. The figures for same-sex pairs include only those who both returned Schoenfeldt's questionnaire on zygosity. Schoenfeldt did not send this questionaire to twins enrolled in different grades. The figures for opposite-sex pairs include 45 pairs enrolled in different grades.

Line 4. These cases were identified by merging records from Talent's 5-year followup file with Schoenfeldt's lists of same-sex pairs with zygosity data and with his full list of mixed-sex pairs. We then eliminated 30 pairs who lacked either a composite test score or data on educational attainment.

were enrolled in the same school, who had the same last name, who listed parents with the same first name, and who returned 5-year followup data.[9] When we found more than two siblings in the same school, we took the two oldest. These rules yielded 867 pairs of siblings with composite test scores and educational attainment data.

We did not search the Talent files for siblings who failed to return 5-year followup data. Thus we cannot give a precise estimate of the amount of sample attrition. The individual response rate in the 5-year followup averaged 35 percent. If responses had been randomly distributed across families, the expected response rate for pairs would be (0.35) (0.35) = 12 percent. But responses are not likely to be randomly distributed across families. Among mixed-sex twins, for example, the individual response rate is also about 35 percent. But if one twin responded, the chance that the other would respond rose from 35 to 60 percent. The response rate for pairs was thus (0.35) (0.60) = 21 percent. The response rate for siblings is probably similar.

With attrition of this magnitude, we must clearly be careful in generalizing to larger populations.

C. Measures of Attainment

Our analyses concentrate on three variables: academic achievement in high school, years of school completed, and career plans. We measured them as follows:

Academic Achievement is a composite score constructed by the Project Talent staff. We selected it over other alternative scores because its correlation with years of school completed was higher than that of any other Talent composite and only slightly lower than the multiple correlation of all available scores with educational attainment (Crouse, forthcoming). An individual's achievement score is the unequally weighted sum of correct answers to 289 items. After weighting, roughly half of these items deal with vocabulary, reading comprehension, and verbal skills. Roughly 40 percent require quantitative skills.[10] The remaining 10 percent deal with "abstract reasoning" and "creativity".

Respondents took the test at different ages. We expected educational aspirations to affect how much work students did in high school, and hence how much they learned. We also expected 12th grade test performance to have more effect on students' chances of attending college than 9th grade performance. We therefore expected a respondent's educational attainment to correlate better with his 12th grade score than with his 9th grade score. Talent's longitudinal data showed, however, that educational plans did not appreciably affect cognitive growth between 9th and 12th grades. Nor did changes in an individual's test performance between 9th and 12th grades affect his later educational attainment. An individual's 9th grade score therefore predicted his educational attainment as accurately as his 12th grade scores.[11] These conclusions also held for cross-sectional comparisons of 9th, 10th, 11th, and 12th graders in the present sample. Such results suggest that the academic achievement composite predicts educational attainment not because it measures actual academic achievement, which changes from 9th to 12th grade, but because it is a proxy for aptitudes that remain stable from 9th to 12th grade. We will therefore treat our achievement scores as fallible measures of academic aptitude. This implies that we should treat changes in students' scores between 9th and 12th grades as measurement errors. We do not know the magnitude of these errors.[12]

Since raw scores increase with age, while aptitude presumably does not, we standardized all scores by age. To do this, we first regressed the raw scores on age. This regression was essentially linear for 14 to 17 year olds, with scores increasing by a sixth of a standard deviation each year. Eighteen and 19 year olds scored below 17 year olds, presumably because the abler 18 and 19 year olds were in college. Thirteen year olds were less than a sixth of a standard deviation below the 14 year olds, presumably because 13 year olds with low scores were still in 8th grade. These results were consistent with the assumption that the regression slope for the population as a whole was linear from 13 to 19, increasing at about 0.17 standard deviations per year. We used this assumption to standardize all scores. This meant that the 13 year olds in our sample had above-average aptitudes, while 18 and 19 year olds had below-average aptitudes. This is what one would expect on the basis of known selection biases. We did not try to adjust the slope for 14-17 year olds to take account of the bias introduced by differential attrition. Since age explained less than three percent of the variance in observed scores, and less than one percent of the variance in adjusted scores, this standardization procedure should be adequate.

Once we had standardized for age, we transformed the resulting aptitude scores so they would have a mean of approximately 100 and a standard deviation of approximately 10 in the full Talent sample. This transformation may not be exact, since Talent only publishes norms for its tests by grade level, not age.

Educational Attainment. The "5-year followup" was actually conducted five years and four months after expected high school graduation. A student who had remained in school and progressed at the usual rate would therefore have been entering his second year of graduate work. If a student was not in school at the time of the survey, we assigned him the highest grade he had completed. If he was enrolled in college as a freshman or sophomore, we assigned him two more years than he had completed. If he was enrolled as a junior or senior, we assumed he would earn a B.A. but no more. If he was enrolled as a part-time graduate student, we assumed he would complete one further year of graduate work. If he was a full-time graduate student, we assumed he would complete two years of graduate study. If he had completed technical or vocational training after high school, we assigned him 13 years of education. These assignments obviously underestimate some people's eventual attainment and overestimate others', but the discrepancies should not be large relative to the total variance. In a small subsample that was resurveyed six years later our estimates correlated 0.91 with educational attainment at 28. The correlation between two independent estimates of educational attainment rarely exceeds this level even when the measures are obtained within a few weeks of one another.

Career Plans. The 5-year followup asked each respondent "What occupation do you plan to make your life work?" Ninety percent of the males and 58 percent of the females answered this question. As a result, 646 of the 1203 pairs with test scores and educational attainment had career plans for both respondents.

Talent coded responses to this question into quite detailed occupational categories. We matched the Talent categories to Census categories as best we could and then assigned them a Duncan score (Duncan, 1961). Duncan's scoring system is based on the characteristics of the males engaged in a given occupation, but it also seems to give a fairly satisfactory estimate of the relative status of females in various occupations. It is probably not appropriate for comparing the status of males and females, since there is considerable sex role differentiation within occupations.

D. Sample Biases

Table 2 shows the means, standard deviations, and correlations among our basic variables for the 1203 pairs of twins and siblings with data on academic achievement and educational attainment. Table 3 shows analogous data for the 646 pairs who both reported career plans. These tables were constructed from a file in which every individual appears twice: once as the first member of a pair and once as the second member. This eliminates random differences between the first and second member of each pair. We weight each pair by 0.50 in order to get the correct N and correct standard errors. We use A to denote academic achievement, B to denote educational attainment, and C to denote career plans. We use a prime to distinguish the second member of each pair from the first. Thus r_{AB} is the correlation between an individual's academic achievement and his educational attainment, while $r_{AB'}$ is the correlation between his academic achievement and his twin or sibling's educational attainment.

Since the samples covered by Tables 2 and 3 were subject to extraordinarily high attrition, we expected them to differ in some respects from the cohort from which they were drawn. Table 2 shows that our sample of twins and siblings scores 0.547 standard deviations above the estimated mean for the full Talent sample on academic achievement. There are no significant differences among the means for MZ twins, DZ twins, and siblings. Most other twin samples exhibit depressed means. All our means are inflated by selective response to the various follow-up questionnaires. Since same-sex twins had to return one more questionnaire (i.e. the one on zygosity) than ordinary siblings to be included in our sample, their mean is more inflated than the sibling mean. After controlling parental status, our MZ twins score about a sixth of a standard deviation below siblings, while our DZ twins score a tenth of a standard deviation below siblings. Both differences are significant, though the difference between MZ and DZ twins is not. Males and females do not differ significantly on this composite, though they do differ significantly on some subtests.

Table 2 also shows that our sample averaged 14.3 years of schooling.

Table 2: Means, Standard Deviations, and Correlations of Academic Achievement (A) and Educational Attainment (B) for Twins and Siblings with Data on Both A and B

Sample	Abbreviation	No. of Pairs	$\bar{A}$	s_A	$\bar{B}$	s_B	r_{AB}	$r_{AA'}$	$r_{BB'}$	$r_{AB'}$
MZ Twins	M	166	104.776	9.631	14.587	1.984	.621	.864	.780	.586
Male Pairs	Mm	76	105.821	9.852	15.211	1.900	.677	.853	.745	.609
Female Pairs	Mf	90	103.894	9.377	14.061	1.903	.573	.871	.770	.564
DZ Twins	D + D"	170	105.362	9.079	14.282	2.007	.589	.567	.521	.472
Same-Sex Pairs	D	77	106.166	9.361	14.364	1.963	.565	.722	.590	.544
Male Pairs	D m	26	105.124	9.879	14.615	1.952	.588	.641	.547	.551
Female Pairs	D f	51	106.696	9.089	14.235	1.966	.571	.767	.605	.558
Mixed-Sex Pairs	D"	93	104.697	8.808	14.215	2.045	.610	.416	.467	.411
Siblings	S + S"	867	105.629	9.398	14.227	1.959	.535	.434	.492	.329
Same-Sex Pairs	S	410	105.693	9.704	14.311	1.984	.540	.451	.540	.350
Male Pairs	Sm	205	105.742	9.783	14.590	1.890	.563	.483	.481	.380
Female Pairs	Sf	205	105.644	9.636	14.032	2.038	.529	.417	.577	.329
Mixed Sex Pairs	S"	457	105.572	9.119	14.151	1.935	.531	.417	.443	.309
Pairs Born 10-18 Months Apart		274	105.640	9.085	14.117	1.854	.488	.445	.412	.273
Pairs Born 19-27 Months Apart		316	105.517	9.386	14.215	2.000	.546	.431	.516	.318
Pairs Born 28-36 Months Apart		189	106.514	9.394	14.524	2.021	.532	.436	.507	.366
Pairs Born 37 or More Months Apart		83	104.419	10.173	14.006	1.946	.608	.398	.541	.412
Total		1203	105.474	9.387	14.284	1.972	.551	.514	.538	.383

Table 2 (cont'd)

Other Twin and Sibling Samples										
Schoenfeldt's Same-	MZ	410	100.9	9.2	NA	NA	NA	.838	NA	NA
Sex Talent Twins[a]	DZ	175	102.1	9.4	NA	NA	NA	.625	NA	NA
NAS-NRC Male	MZ	1022	NA	NA	13.5	3.0	NA	NA	.765	NA
Twins born 1917-1927[b]	DZ	914	NA	NA	13.3	3.1	NA	NA	.545	NA
NORC Brothers Born 1910-1949[c]		150	NA	NA	12.447	3.168	NA	NA	.528	NA
Kalamazoo Brothers born 1917-1938[d]		346	100.893	9.579	13.197	2.730	.576	.469	.549	.400
Individuals in This Sample										
Males		1164	105.179	9.688	14.543	1.937	.586			
Females		1242	105.750	9.091	14.042	1.975	.536			
Individuals in Other Samples										
CPS 25-29 in 1971[e]		10,000			12.366	2.778				
CPS 25-29, B ≥ 9 in 1971		9,200			12.871	2.205				

[a]Our tabulations of Schoenfeldt's data

[b]Behrman, Taubman and Wales, Chapter 3 of this volume.

[c]Eaglesfield (forthcoming)

[d]Olneck (1976); s_A is transformed to our metric (i.e. s_A = 10 for the full population) assuming that the population variance for Otis and Terman IQ's was 16.

[e]U.S. Bureau of the Census, Current Population Reports, Population Characteristics, Series P-20, No. 229, "Educational Attainment: March 1971," Washington, D.C. 1971, Table 1. The N's are approximate.

Table 3: Means, Standard Deviations, and Correlations of Academic Achievement (A), Educational Attainment (B), and Career Plans (C) for Twins and Siblings with Data on A, B, and C.

Sample		# of Pairs	$\bar{A}$	s_A	$\bar{B}$	s_B	$\bar{C}$	s_C	r_{AB}	r_{AC}	r_{BC}	$r_{AA'}$	$r_{BB'}$	$r_{CC'}$	$r_{AB'}$	$r_{AC'}$	$r_{BC'}$
MZ Twins		99	106.441	9.519	15.141	1.893	63.056	19.082	.594	.555	.650	.842	.756	.584	.528	.528	.528
Male Pairs		65	105.954	9.870	15.292	1.898	65.254	20.493	.664	.638	.642	.836	.712	.617	.581	.614	.537
Female Pairs		34	107.373	8.806	14.853	1.863	58.853	15.322	.483	.399	.670	.852	.837	.421	.453	.365	.488
DZ Twins		81	107.372	9.161	14.982	1.938	62.753	17.671	.654	.445	.534	.568	.535	.146	.530	.293	.414
Same-Sex Pairs		40	108.295	8.953	15.013	1.832	64.113	18.419	.595	.484	.572	.639	.528	.127	.565	.228	.326
Male Pairs		23	106.536	9.541	14.674	1.910	62.652	22.105	.633	.559	.636	.590	.494	.174	.600	.288	.349
Female Pairs		17	110.674	7.589	15.471	1.637	66.088	11.784	.455	.245	.414	.679	.525	-.141	.429	-.014	.240
Mixed-Sex Pairs		41	106.471	9.326	14.951	2.048	61.427	16.916	.708	.399	.504	.496	.541	.158	.503	.350	.502
Siblings		466	106.982	9.452	14.873	1.906	60.851	20.627	.524	.448	.568	.430	.404	.215	.284	.233	.278
Same-Sex Pairs		247	106.770	9.774	14.879	1.893	61.668	21.098	.534	.472	.570	.446	.442	.264	.304	.265	.296
Male Pairs		173	106.082	9.911	14.697	1.888	60.861	23.233	.560	.523	.571	.474	.469	.280	.376	.333	.337
Female Pairs		74	108.378	9.282	15.304	1.843	63.554	14.861	.441	.286	.596	.348	.333	.153	.083	-.009	.138
Mixed-Sex Pairs		219	107.220	9.079	14.868	1.922	59.929	20.066	.514	.421	.568	.408	.363	.150	.260	.196	.258
Total		646	106.948	9.422	14.928	1.909	61.427	20.055	.549	.462	.576	.511	.476	.261	.350	.282	.330
<u>Other Twin and Sibling Samples</u>																	
NAS-NRC Male	MZ	1022	NA	NA	13.6	3.03	54.3	23.1	NA	NA	.495	NA	.765	.389	NA	NA	.392
Twins (40-49)[a]	DZ	914	NA	NA	13.4	3.12	53.8	22.6	NA	NA	.458	NA	.545	.163	NA	NA	.266
NLS Brothers (14-24)[b]		580	101.8	9.9	14.8	2.3	8.67	.404	.482	.275	.415	.555	.508	.109	.333	.204	.228
Kalamazoo Brothers (35-54)[c]		346	100.893	9.579	13.197	2.730	49.912	23.157	.576	.453	.591	.469	.549	.309	.400	.300	.378
NORC Brothers (25-64)[d]		150	NA	NA	12.447	3.168	41.093	24.896	NA	NA	.595	NA	.528	.371	NA	NA	.401

Table 3 (cont'd)

Individuals in This Sample																
Males	782	105.987	9.778	14.839	1.935	62.032	22.230	.589	.531	.588						
Females	510	108.422	8.653	15.065	1.862	60.500	16.138	.471	.334	.578						
OCG Non-Negro Men[e] (25-64)	7000	NA	NA	11.15	3.40	39.6	24.5	NA	NA	.606	NA	.573	NA	NA	NA	.401
(25-34); non-farm)	1500	NA	NA	12.38	3.04	43.3	25.0	NA	NA	.651	NA	.536	NA	NA	NA	.391

[a]Behrman, Taubman and Wales, Chapter 3 of this volume.
C = Duncan score at age 40-49.

[b]Griliches (1975) and in correspondence.
A = Miscellaneous school tests standardized to our metric.
B = *Expected* eventual educational attainment.
C = Mean earnings of *expected* occupation.

[c]Olneck (1976). See Notes to Table 2.
C = Duncan score of occupation at age 35-54.

[d]Eaglesfield (forthcoming).
C = Duncan score of occupation at age 25-64.

[e]Duncan, Featherman, and Duncan (1972:263). Pairwise correlations for non-Negro men aged 25-64 in the civilian labor force in 1962, half of whom had a brother and reported his education.
C = Duncan score of occupation at age 25-64.

The CPS cohort born at the same time averaged 12.4 years of school when it was 25-29. Roughly 0.5 years of this difference is due to the fact that Talent's original sample excluded individuals with less than 9 years of schooling. Another 0.2 years is due to the fact that our education measure includes vocational training whereas the CPS measure excludes such training. The remaining 1.2 year difference is largely attributable to the fact that those with 9 to12 years of schooling were much less likely to return Talent's 5-year followup questionnaire. The differences between means for the MZ twins, DZ twins, and siblings are insignificant. Males get significantly more education then females.

Table 3 shows that our respondents planned careers in occupations whose Duncan scores average 61. This is 20 points above the observed mean for 25-64 year old workers in 1973. The differences between means for MZ twins, DZ twins and siblings are insignificant as is the difference between males and females.

Turning from means to standard deviations, Table 2 shows that the standard deviation of academic achievement is only slightly smaller in our sample than in the full Talent sample (where it is roughly 10). The standard deviation of educational attainment is no smaller than in the representative Talent sample, though it is somewhat restricted relative to the overall birth cohort. The standard deviations of test scores and educational attainment do not differ significantly between MZ twins, DZ twins and siblings. In Table 3, the standard deviation of career plans is significantly lower for DZ twins than for siblings. But this is the only significant difference in the 18 possible comparisons of means and standard deviations for MZ twins, DZ twins, and siblings. Since this difference has a significance level between 0.05 and 0.01, it too is probably due to chance. The test score and career plans variances for males are significantly higher than for females. These differences are consistent with other studies. The "surprise" is that the variance of educational attainment is as great for females as for males.

Changes in the means and variances may or may not affect the correlations among outcome measures. Tables 2 and 3 show that where analogous correlations are available from other samples, they are quite similar to ours. Since some of our comparison samples come from quite different periods, and since the variances have clearly changed over time, this is somewhat surprising. This result, which is quite common in stratification research, suggests that the size of an individual's relative advantage (e.g. 1.0 vs 0.5 standard deviations) is more important than the absolute size of a given advantage. If this is the case, standardizing all variables to unit variances will yield results of greater generality than using unstandardized variables.

One final problem deserves attention. Schoedfeldt's validity study indicated that something like 9.6 percent of twins diagnosed as MZ were probably DZ. 17.1 percent of those diagnosed as DZ were probably MZ. The differences between MZ and DZ pairs in Tables 2 and 3 are therefore likely to be smaller than the differences between correctly diagnosed pairs. To estimate the magnitude of this error, we designate the observed intraclass correlations between twins' phenotypic values as $r_{PP'\underline{M}*}$ and

$r_{PP'\underline{D}*}$. The true correlations are $r_{PP'\underline{M}}$ and $r_{PP'\underline{D}}$. If $E_{\underline{M}*}$ is the percentage of same-sex twins diagnosed as MZ who are really DZ, and $E_{\underline{D}*}$ is the percentage of same-sex twins diagnosed as DZ who are really MZ, and if the observed means and variances for MZ and DZ pairs are the same, one can show that:

(11) $$r_{PP'\underline{M}*} = (1 - E_{\underline{M}*})\, r_{PP'\underline{M}} + E_{\underline{M}*}\, r_{PP'\underline{D}}$$

and

(12) $$r_{PP'\underline{D}*} = (1 - E_{\underline{D}*})\, r_{PP'\underline{D}} + E_{\underline{D}*}\, r_{PP'\underline{M}}$$

Solving for the true correlations we get:

(13) $$r_{PP'\underline{M}} = r_{PP'\underline{M}*} + \frac{E_{\underline{M}*}\,(r_{PP'\underline{M}*} - r_{PP'\underline{D}*})}{1 - E_{\underline{M}*} - E_{\underline{D}*}}$$

and

(14) $$r_{PP'\underline{D}} = r_{PP'\underline{D}*} - \frac{E_{\underline{D}*}\,(r_{PP'M*} - r_{PP'D*})}{1 - E_{\underline{M}*} - E_{\underline{D}*}}$$

Using Schoenfeldt's point estimates for E_{M*} and E_{D*} (i.e. 0.096 and 0.171) and solving equations 13 and 14 simultaneously we obtain:

$$\hat{r}_{PP'\underline{M}} = r_{PP'\underline{M}*} + 0.131\,(r_{PP'\underline{M}*} - r_{PP'\underline{D}*}) =$$

$$1.131\, r_{PP'\underline{M}*} - 0.131\, r_{PP'\underline{D}*}$$

and

$$\hat{r}_{PP'\underline{D}} = r_{PP'\underline{D}*} - 0.233\,(r_{PP'\underline{M}*} - r_{PP'\underline{D}*}) =$$

$$1.233\, r_{PP'\underline{D}*} - 0.233\, r_{PP'\underline{D}*}$$

These formulas imply that accurate diagnosis would not alter the correlations in Table 2 dramatically, but that the effects are large enough to justify retaining the correction in future computations. Since Taubman reports that his diagnosis procedure is 95 percent accurate, the corrections for his results would be smaller (i.e. $r_{PP'M} \sim 1.056\, r_{PP'\underline{M}*} - 0.056\, r_{PP'\underline{D}*}$ and $r_{PP'D} \sim 1.056\, r_{PP'\underline{D}*} - 0.056\, r_{PP'\underline{M}*}$), and could probably be neglected entirely with little loss of information. These corrections are not, of course, needed for mixed-sex twins.

III. Assumptions Required to Estimate Genotypic Variance for Single Traits

This section develops a formula for estimating the genotypic variance of test scores, educational attainment, and career plans from the data in Tables 2 and 3. The formula depends on five assumptions about twins. The bulk of the discussion focuses on the empirical evidence for and against these assumptions.

We begin by scaling all variables as deviations from the population mean. The unstandardized equation implied by Figure 2 is still:

$$(4) \quad \underline{P_i} = h\underline{G_i} + i\underline{X_i} + j\underline{U_i}.$$

But G is the mean phenotype of individuals who start life randomly distributed across a specified range of environments. This means that a unit increase in G will produce a unit increase in P once X is controlled. Thus h = 1. We can scale exogenous environmental influences in the same way by regressing P on all the different exogenous environmental variables (X_1, X_2...X_n) that affect P, while holding G constant. We then set X equal to $B_1X_1 + B_2X_2...B_nX_n$. This will make i = 1 in equation 4. Finally, we scale the interaction term (U) so as to make j = 1. Equation 4 then reduces to:

$$(15) \quad \underline{P_i} = \underline{G_i} + \underline{X_i} + \underline{U_i}$$

The character of U deserves some comment. U represents an error term, attributable to interactions between genotype and environment. These interactions may be either simple or complex. What we call "simple" interactions arise when the effects of G and X are not additive. If, for example, P = GX, an additive equation like 15 will not yield precise predictions of P. One can often detect interactions of this kind by looking for heteroscedasticity. Jinks and Fulker (1970) suggest, for example, that if the true relationship takes the form P = GX, the absolute difference between identical twins' phenotypes ($P_i - P_i'$) will tend to increase as pairs' mean phenotypes ($\bar{P}_i$) increase. This test is particularly strong for identical twins reared apart. One can often eliminate simple interactions of this kind by rescaling P. If P = GX, for example, log P = log G + log X. Thus if we convert P to logarithms, we can scale G and X so as to make their effects perfectly additive.

What we call "complex" interactions are more difficult to detect and eliminate, since they involve the <u>components</u> of G and X rather than the overall index. Suppose, for example, that height is completely determined by genes (which it is not) and that the ability of a student's high school basketball coach is a completely exogenous environmental variable. Suppose that extremely tall high school students enjoy a slight edge in getting basketball scholarships, but that extremely tall non-basketball players are regarded as slightly odd, so that on the average extremely tall students get no more education than anyone else. Likewise, suppose that having a talented high school basketball coach slightly increases the percentage of students from a school who get basketball scholarships, but that a talented coach also reduces academic effort in his school so that on the average students from such schools have the same probability of attending college as students from other schools. Under these circumstances height has no effect on G, and the ability of the high school basketball coach has no effect on X. A tall student in a school with a talented basketball coach will therefore be at the mean on both G and X, and will have the same predicted educational attainment as anyone else. In fact, however, a tall student with a talented coach is much more likely than anyone else to get a basketball scholarship. His expected value of U is therefore positive, at least for educational attainment. Such an interaction would not necessarily signal its presence by making $P_i - P'_i$ correlate with $\bar{P}_i$. The absence of such a correlation does not therefore prove the absence of complex interactions. The only way to detect such interactions is actually to measure the specific genetic and environmental variables that affect P. Nor can we eliminate complex interactions by rescaling P. Transforming P can eliminate errors due to misspecification of the relationship between G, X, and P, but it cannot eliminate errors that arise because P varies among individuals with identical values on both G and X. The only way to eliminate such errors would be to decompose both G and X into their components and estimate the additive and non-additive effects of these components on P.

For practical purposes, then, it is more useful to think of U as a random error term, uncorrelated with G and X by construction, than to think of it as a conventional interaction term.

We now introduce six new measures, defined as follows:

$\hat{G}$ is the predicted mean genotype of all children with a given set of parents. If parents had an infinite number of children, $\hat{G}$ would equal the mean of G for all their children ($\bar{G}_C$). Since parents do not have an infinite number of children, $\bar{G}_C$ will differ in random ways from $\hat{G}$. The standard deviation of $\bar{G}_C - \hat{G}$ will be equal to $1/\sqrt{N}$, where N is the number of children. If all genetic effects were additive, $\hat{G}$ would equal the mean of the parents' genotypes ($\bar{G}_P$). Since genetic effects are not completely additive, i.e., since there is some dominance and epistasis, $\hat{G}$ is not exactly equal to $\bar{G}_P$.

$\tilde{G}$ = the genotypic deviation of any particular individual from the expected genotypic value for his or her family. For the first member of the ith pair of twins, $\tilde{G}'_i = G_i - \hat{G}_i$ while for the second member $\tilde{G}_i = G'_i - \hat{G}_i$. The expected value of $\tilde{G}'_i$ is zero, but the mean value for a given pair need not be zero, since both members of the pair may be either above or below the family's expected mean.

$\hat{X}$ is the predicted mean of the exogenous environmental variables that affect all individuals in a given family, when these variables are scaled in the manner described above. Once again, if parents had an infinite number of children, $\hat{X}$ would equal $\bar{X}_C$.

$\tilde{X}$ = the deviation of the exogenous environmental influences on a given individual from the expected value for all individuals in his family. This definition implies that $\tilde{X}_i = X_i - \hat{X}_1$ and $\tilde{X}'_i = X'_i - \hat{X}_i$.

$\hat{U}$ is the difference between $\hat{P}$ and the value of $\hat{P}$ predicted using an additive equation with $\hat{G}$ and $\hat{X}$ as independent variables. $\hat{P}$ denotes the expected phenotypic value for all individuals in a given family. If we know $\hat{G}$, $\hat{X}$, and $\hat{U}$, we can predict $\hat{P}$ without error. If we know only $\hat{G}$ and $\hat{X}$, we will make erroneous predictions of $\hat{P}$. We denote these errors as $\hat{U}$. It follows from this definition that $\hat{U}$ is uncorrelated with $\hat{G}$ and $\hat{X}$.

$\tilde{U}$ = the difference between P and the value predicted on the basis of $\hat{G}$, $\tilde{G}$, $\hat{X}$, $\tilde{X}$, and $\hat{U}$. Note that this residual can involve not only interactions between genetic and environmental deviations from the family's expected values (i.e., between $\tilde{G}$ and $\tilde{X}$), but interactions between the components of $\tilde{G}$ and $\hat{X}$ or between the components of $\tilde{X}$ and $\hat{G}$. Suppose, for example, that all members of a given family attend a school with a good basketball coach, but that only the tallest member gets a scholarship as a result. Genetic variations within a family then interact with the family's common environment.

We now write a modified version of equation 15 in which:

$$(16) \quad P_i = B_{\hat{G}}\hat{G}_i + B_{\tilde{G}}\tilde{G}_i + B_{\hat{X}}\hat{X}_i + B_{\tilde{X}}\tilde{X}_i + B_{\hat{U}}\hat{U}_i + B_{\tilde{U}}\tilde{U}_i$$

Since $\hat{G}_i + \tilde{G}_i + G_i$ and $\hat{X}_i + \tilde{X}_{iU} = X_i$, and since $P_i = G_i + X_i + U_i$, it might seem that $B_{\hat{G}} = B_{\tilde{G}} = h = 1$, and $B_{\hat{X}} = B_{\tilde{X}} = i = 1$, and $B_{\hat{U}} = B_{\tilde{U}} = j=1$. This need not be so, however.

Suppose that parents allocate their time, energy, and money in somewhat Rawlsian fashion, trying to redress the genetic disadvantages of their least favored children. Suppose that they do this to a significantly greater extent than other institutions, such as schools. A genetic difference between two children from the same home will then produce a smaller phenotypic difference between these two children than would an identical genetic difference between the means for two different homes. This will make $B_{\tilde{G}} < B_{\hat{G}}$. In our earlier model, such phenomena would have implied an interaction between the components of G and the components of X. To see why this is so, consider the case of a genetically disadvantaged student. If the respondent has genetically advantaged siblings, the parents will be able to devote a lot of resources to the respondent. This may mute the effects of the genetic disadvantage. If the child has siblings even more disadvantaged than himself, the parents will devote their energies to these siblings. The respondent will be worse off as a result. From the respondent's viewpoint, then, sibling genotypes are part of the exogenous environment. Advantaged siblings help. Disadvantaged siblings hurt. The value of a unit increase in G depends to some extent on this exogenous environmental factor. Equation 15 embodies this interaction in U. Equation 16 captures it by letting $B_{\hat{G}}$ differ from $B_{\tilde{G}}$. This makes $s^2_{\hat{U}} + s^2_{\tilde{U}}$ less than s^2_U.

The same logic applies to exogenous environmental factors. If one child suffers from a debilitating disease, for example, the parents may be able to compensate the child with extra attention, special schooling, and the like. If every child in the family suffers from the disease because of unhealthy living conditions, the parents will be less able to offset its effects. Thus $B_{\hat{X}}$ may differ from $B_{\tilde{X}}$.

Both examples assume that families do more than the larger society to offset the effects of genetic or exogenous environmental disadvantages. So long as this is true, $B_{\hat{G}} > B_{\tilde{G}}$ and $B_{\hat{X}} > B_{\tilde{X}}$. But one can imagine societies that try to offset the effects of coming from a disadvantaged family, while families make less effort to offset the effects of being the least advantaged sibling. This could happen if, for example, schools allocated extra effort to low SES children, regardless of their actual abilities, while parents favored their ablest children. This could end up making $B_{\hat{G}} < B_{\tilde{G}}$ or $B_{\hat{X}} < B_{\tilde{X}}$.

Now let us consider the likely correlations among the variables on the right side of equation 16. If parents are genetically advantaged for some particular phenotype, $\hat{G}$ will be positve for their children as well. Such parents are also likely to provide their children with unusually favorable environments, regardless of the child's genotype. This will make $\hat{X}$ positive. It follows that $r_{\hat{G}\hat{X}}$ will be positive.[13]

Unlike $\hat{G}$, $\tilde{G}$ is initially random.[14] Genetic deviations from the expected family mean can end up correlated with environmental factors, but only if genotype affects environment. Environmental responses to genotype are by definition endogenous. Exogenous environmental factors cannot, then, be correlated with genetic deviations from the expected family mean. This makes $r_{\tilde{G}\tilde{X}} = 0$.

$\hat{G}$, $\hat{X}$ and $\hat{U}$ only vary between families. $\tilde{G}$, $\tilde{X}$, and $\tilde{U}$ can also vary between families, but the variations are by definition random. This ensures that $r_{\hat{G}\tilde{G}} = r_{\hat{G}\tilde{X}} = r_{\hat{G}\tilde{U}} = r_{\hat{X}\tilde{G}} = r_{\hat{X}\tilde{X}} = r_{\hat{X}\tilde{U}} = r_{\hat{U}\tilde{G}} = r_{\hat{U}\tilde{X}} = r_{\hat{U}\tilde{U}} = 0$. Furthermore, $r_{\hat{G}\hat{U}} = r_{\hat{E}\hat{U}} = r_{\tilde{G}\tilde{U}} = r_{\tilde{X}\tilde{U}} = 0$ by construction. Taking variances from equation 16 therefore yields:

$$(17)\quad s^2_P = B^2_{\hat{G}} s^2_{\hat{G}} + B^2_{\tilde{G}} s^2_{\tilde{G}} + B^2_{\hat{X}} s^2_{\hat{X}} + B^2_{\tilde{X}} s^2_{\tilde{X}} + B^2_{\hat{U}} s^2_{\hat{U}} + B^2_{\tilde{U}} s^2_{\tilde{U}} + 2B_{\hat{G}} B_{\hat{X}} s_{\hat{G}} s_{\hat{X}} r_{\hat{G}\hat{X}}$$

We will call $s^2_{\hat{G}}$ the between-family genetic variance, $s^2_{\tilde{G}}$ the within-family genetic variance, $s^2_{\hat{X}}$ the between-family exogenous environmental variance, $s^2_{\tilde{X}}$ the within-family exogenous environmental variance, $s^2_{\hat{U}}$ the between-family interaction variance, and $s^2_{\tilde{U}}$ the within-family interaction variance.

In order to make heritability estimates for the general population, one must assume that the standard deviations and regression coefficients in equations 16 and 17 are the same for both MZ and DZ twins as for the general population. This statistical assumption depends on three substantive assumptions.

I. Twins are born with the same distribution of genotypes and encounter the same range of environments as other children.

II. Twins do not influence one another.

III. Having a twin does not alter the effects of one's genes or one's environment on one's phenotype.

We will discuss these three assumptions in turn.

Assumption I. (Representativeness of Twins). Bulmer (1970) reports that MZ twinning is an essentially random phenomenon, with no demographic correlates. It follows that the variance of $\hat{G}$, $\tilde{G}$, $\hat{X}$, $\tilde{X}$, $\hat{U}$, and $\tilde{U}$ should be the same for MZ twins as for the general population.

Bulmer reports that DZ twinning is more common among blacks than among whites, and more common among whites than among Orientals. Since these ethnic groups have somewhat different gene pools (e.g., for skin color and facial features), phenotypic differences in their test performance and educational attainment are by definition traceable to genotypic differences. (As noted in Section 1, this does not mean that the three groups' phenotypes would differ if all three groups encountered the same range of environments. Here as elsewhere, genes may affect test scores and education by affecting the environment.) Since ethnic groups differ with respect to both mean genotype and DZ twinning rates, DZ twins' mean genotype for these traits must differ from the general population's mean genotype. The genotypic variance among DZ twins may also differ from the genotypic variance in the general population.

Bulmer also reports that DZ twinning is more common among women with large numbers of previous pregnancies. This means that DZ twins are more likely to be born into large families and are likely to be younger children. Since family size has substantial effects on both test scores and educational attainment, and since birth order also has modest effects on both these outcomes, the mean value of the exogenous environmental factors influencing test scores and educational attainment is presumably lower for DZ twins than for the general population. The environmental variance may also be restricted among DZ twins.

These biases do not seem to affect our particular sample, however. Our sample contains virtually no blacks or Orientals, and the mean family size for DZ twins does not differ significantly from the mean for MZ twins. (Mean family size for twins exceeds the mean for ordinary siblings by about one.) The phenotypic means and variances for test scores, education, and career plans do not differ significantly for MZ twins, DZ twins, and ordinary siblings. In what follows we will therefore assume that our twin sample does not differ from our full sample in either genetic or environmental terms, though our full sample clearly differs from the population as a whole in these respects.

If we use $\underline{M}$ to denote MZ twins, $\underline{D}$ to denote DZ twins, and no subscript to denote the full sample, our argument implies that:

$$(18)\quad s^2_{\hat{G}\underline{M}} = s^2_{\hat{G}\underline{D}} = s^2_{\hat{G}}$$

$$(19)\quad s^2_{\tilde{G}\underline{M}} = s^2_{\tilde{G}\underline{D}} = s^2_{\tilde{G}}$$

$$(20)\quad s^2_{\hat{X}\underline{M}} = s^2_{\hat{X}\underline{D}} = s^2_{\hat{X}}$$

$$(21)\quad s^2_{\tilde{X}\underline{M}} = s^2_{\tilde{X}\underline{D}} = s^2_{\tilde{X}}$$

$$(22)\quad s^2_{\hat{U}\underline{M}} = s^2_{\hat{U}\underline{D}} = s^2_{\hat{U}}$$

$$(23)\quad s^2_{\tilde{U}\underline{M}} = s^2_{\tilde{U}\underline{D}} = s^2_{\tilde{U}}$$

$$(24)\quad r_{\hat{G}\hat{X}\underline{M}} = r_{\hat{G}\hat{X}\underline{D}} = r_{\hat{G}\hat{X}}$$

Assumption II (Absence of Reciprocal Influences). Suppose friends emulate one another. If this were the case, a respondent's test performance and education would depend partly on his choice of friends. One would expect most genetically advantaged respondents to choose genetically advantaged friends, while genetically disadvantaged respondents would usually choose genetically disadvantaged friends. Suppose that best friends' genotypes correlate 0.25 for the population as a whole, but suppose identical twins almost always choose each other as best friends. For them, then, the correlation between respondent's genotype and best friend's genotype approaches 1.00. Now consider two individuals who both rank one standard deviation above the mean in terms of test score genotype, one of whom has an identical twin while the other does not. Suppose the phenotypic standard deviation is 10 points, as it is for our test, and that the genotypic standard deviation is 5 points. Other things equal, an individual whose genotype ranked one standard deviation above the mean would have a predicted phenotypic score of 100 + 5 = 105 points. This expected value assumes that the respondent's best friend will have a genotype 0.25 standard deviations above the mean, since that is the usual value for respondents with genotypes a full standard deviation above the mean. But if the respondent has an identical twin, and if his twin is his best friend, his best friend will have a genotype a full standard deviation above the mean. The respondent's predicted test scores will therefore exceed 105. The implied values of $B_{\hat{G}}$ and $B_{\tilde{G}}$ will thus be greater for MZ twins than for ordinary siblings. (Alternatively, we could constrain $B_{\hat{G}}$ and $B_{\tilde{G}}$ to be the same for MZ twins as for others. Then $\tilde{U}$ would correlate with $\tilde{G}$ among MZ twins.)

This logic may or may not apply to fraternal twins. The correlation between their genotypes is only about 0.5. Thus even if DZ twins are as likely as MZ twins to choose one another as best friends, the genotypic correlation between the respondent and his best friend will be lower for DZ twins than for MZ twins. Indeed, if assortative friendship normally produced a genotypic correlation in excess of 0.5 between best friends, genotypes could end up having less effect on DZ twins' phenotypes than on ordinary siblings' phenotypes. This seems unlikely, but it underlines the point that the values of $B_{\hat{G}}$ and $B_{\tilde{G}}$ may be about the same for DZ twins as for ordinary siblings.

In order to estimate the effect of each twin on the other, we need to identify one or more instrumental variables, i.e., respondent characteristics that can affect a respondent's own phenotype but can only affect the twin's phenotype indirectly, through the effect of the respondent's own phenotype on his twin's phenotype. Suppose, for example, that we could measure G and G' directly, or at least could measure a significant number of the genetic characteristics that determine G and G'. Suppose, too, that we could convince ourselves that these genetic traits had no visible manifestations other than their effects on a respondent's phenotype. If we then regressed the respondent's phenotype on both G and G' simultaneously, and if we found that G' affected P even with G controlled, we could impute the effect of G' on P to the fact that P' affected P. We could then use the ratio of the coefficient of G' to the coefficient of G to estimate the effect of P' on P. Suppose, for example, that:

(25) $P = B_1 G + B_2 P' + e$

(25') $P' = B_1 G' + B_2 P + e'$

where G is an "invisible" genetic determinant of P, and e and e' are random error terms. The reduced form equation for P is then:

(26) $P = B_1 G + B_1 B_2 G' + B_2^2 P + B_2 e' + e$

$$= \frac{B_1}{1 - B_2^2} G + \frac{B_1 B_2}{1 - B_2^2} G' + \frac{B_2}{1 - B_2^2} e' + \frac{1}{1 - B_2^2} e$$

Dividing the observed coefficient of G' by the coefficient of G would thus give us an estimate of the effect of one twin's phenotype on the other twin's phenotype with genotype controlled, i.e. B_2.

Unfortunately, our data provide only one measure that is clearly a proxy for genotype, namely sex. Furthermore, mean test performance and career plans do not differ by sex, so sex is not a proxy for genotype in these instances. But males do get 0.5 years more education than females in our sample. It follows that $\bar{G}$ for males exceeds $\bar{G}$ for females by 0.5. If twins had no effect on one another, males should get 0.5 years more education than females regardless of whether they had a male or female twin. This is not the case. Males with a male twin get substantially more education than males with a female twin. This is particularly true if the male twin is identical. In order to sort out these effects, we regressed the respondent's educational attainment on nine dichotemous variables, defined as follows:

m = 1 if respondent is male, otherwise 0

m' = 1 if respondent's twin or sibling is male, otherwise 0

m*m' = 1 if m = 1 and m' = 1, otherwise 0

DZ = 1 if respondent has a DZ twin, otherwise 0

MZ = 1 if respondent has an MZ twin, otherwise 0

m*DZ = 1 if DZ = 1 and m = 1, otherwise 0

m'*DZ = 1 if DZ = 1 and m' = 1, otherwise 0

m*m'*DZ = 1 if DZ = 1 and m*m' = 1, otherwise 0

m*MZ = 1 if MZ = 1 and m = 1, otherwise 0

If we enter all nine variables, we get an exhaustive analysis of variance for the main effects and interactions of respondent's sex, his twin or sibling's sex, and the type of pair (MZ, DZ, or ordinary sibling).[15] Only three of the nine variables have significant coefficients. None of the other five coefficients even approach significance. Indeed all five were less than their standard errors. The unstandardized equation (with standard errors in parentheses) was:

$$\text{Education} = \underset{(0.12)}{0.27m} + \underset{(0.10)}{0.28m{*}m'} + \underset{(0.18)}{0.62m{*}MZ} + 14.04$$

The intercept of this regression represents the attainment of females with a female sibling. Males generally get 0.27 years more education than females. Having a male sibling or DZ twin raises male educational attainment by another 0.28 years. But having a male sibling or DZ twin does not have a significant effect on female educational attainment, since m' and m'*DZ are insignificant. This suggests that same-sex pairs may affect one another's educational attainment, while mixed-sex pairs may not. This interpretation is reinforced by the fact that $r_{BB'}$ is consistently (though not significantly) higher for same-sex twins and siblings than for mixed-sex twins and siblings (see Table 2).

If a respondent's sex affected the education of his or her twin or sibling only by affecting the respondent's own education, which then affected the twin or sibling's education, being male would always confer a larger advantage on the respondent than having a male twin or sibling. Yet the point estimate of the coefficient of m*m is slightly larger than the point estimate of the coefficient of m itself (0.28 vs 0.27). This implies that a male gains slightly more from having a male DZ twin or sibling than he gains from being male himself. Substituting into equation 26, the implied value of B_2 is 0.28/0.27 = 1.04, for DZ twins and siblings. The 95% confidence interval for the effect of a one year increase in the education of one's DZ twin or sibling runs from 0.30 to 1.78. Values in excess of 1.00 are logically unacceptable. But a value between 0.30 and 1.00 is quite feasible.

The effect of having a male DZ twin is no greater than the effect of having a male sibling, since m*m'*DZ is insignificant. Since all respondents with an MZ male twin are themselves MZ males, the effect of having an MZ male twin is indistinguishable from the effect of being an MZ male twin oneself. Conceptually, however, we can distinguish the effects by imagining a sample composed of males whose twin died at birth. Would we expect men who had a deceased MZ twin to get more or less education than men who had a deceased DZ twin? Unless the men with MZ twins had different prenatal experiences, were born into different families, or had different response rates, we know of no reason to expect a difference. Since controlling parental SES does not appreciably alter the coefficient of m*MZ, we doubt that the advantage enjoyed by MZ males can be explained by their family characteristics. It could be explained by differential response rates, but there is no direct evidence that education affects MZ response rates more than DZ response rates. We therefore tentatively interpret the positive coefficient of m*MZ as evidence that having a live male MZ twin increases a respondent's educational attainment by 0.62 $\pm$ 0.18 years, over and above the effect of being a male oneself (0.27) and having a male twin or sibling (0.28). If the effect of having an MZ male twin operated only through m' affecting P' and P' affecting P, the implied coefficient of P' would be (0.28 + 0.62)/0.27 = 3.33 $\pm$ 0.76. This is logically impossible, so we must conclude that the model is defective. Either there is some sort of special sampling bias for MZ male pairs, or else having an MZ twin affects a male's educational attainment in some way that is independent of the fact that one's male twin is likely to get more education than average.

In order to make reliable estimates of twins' effects on one another, we would need a much wider range of genetic markers, including some whose effects are not as immediately visible as sex is. We would then have to identify several markers that predicted a respondent's educational attainment. This would give us several different estimates of B_2 in equation 15. If these estimates averaged less than 1.00 and did not differ significantly, we might have some confidence in them

One could conduct a similar analysis using environmental measures that affect a given phenotype and vary within families, so long as they are clearly exogenous. Suppose, for example, that certain respondents had suffered from diseases that affected their eventual educational attainment, while their twins had not had the disease. One could then conduct a regression analysis aimed at determining how much impact the respondent's illness had had on his twin's educational attainment. If we were sure that the respondent's illness had had no direct effect on his twin, we could use its apparent effect to estimate the influence of the respondent's education on his twin's education.

Reciprocal influence invalidates the conventional formula for estimating heritability. If we knew the magnitude of the reciprocal influence, we could derive an alternative formula. Since we do not know the magnitude of such influences, we must make some arbitrary assumption. The simplest assumption is that the effects of sex on educational attainment are atypical, and that twins do not generally affect one another at all. We do not believe this extreme assumption to be correct, but neither do we have any basis for assuming that reciprocal influences are large. Loehlin and Nichols (1976) found no evidence that reciprocal influences had large effects on test performance, but they did not investigate educational attainment.

Assumption III (Having a twin does not alter the effects of either genotype or environment). Even if twins do not affect one another, one could imagine situations in which having a twin would alter the effects of one's own genes or environment. Suppose, for example, that twins spend most of their time with one another, and that this diminishes the effect of parental characteristics. But suppose that twins are so ambivalent about one another that their tendency to differentiate themselves from one another exactly offsets the effects of mutual emulation. The net effect of one's twin's phenotype on the other twin's phenotype will then be zero. But the coefficients of exogenous environmental variables, such as parental test scores and education, will still be reduced.

In order to test the hypothesis that having a twin has no effect on $B_{\hat{G}}$, $B_{\tilde{G}}$, $B_{\hat{X}}$, $B_{\tilde{X}}$, $B_{\hat{U}}$ or $B_{\tilde{U}}$, one must somehow measure $\hat{G}$, $\tilde{G}$, $\hat{X}$, $\tilde{X}$, $\hat{U}$, or $\tilde{U}$. With the exception of sex, we have no genetic measures that affect the phenotypes that interest us. In the case of sex, having a DZ twin does not alter the effects of sex. Having an MZ twin seems to alter the effects of sex, but this may be a matter of reciprocal influence. We therefore turn to the effects of exogenous environmental characteristics ($B_{\hat{X}}$ and $B_{\tilde{X}}$). We examined the coefficients of father's education, father's occupation, and number of siblings. The coefficients of these measures did not differ significantly between MZ twins, DZ twins, and ordinary siblings for any of our three outcomes. The coefficient of test score was significantly higher in the education equation for MZ twins than for DZ twins and siblings combined. But since we tested a total of 24 interaction terms, and since the IQxMZ interaction was not significant at the 0.01 level, we do not take it very seriously. We will therefore proceed on the assumption that having a twin does not alter the effects of genotype or exogenous environmental influences.

If Assumption II and Assumption III were both correct, we could assume that:

$$(27)\ B_{\hat{G}\underline{M}} = B_{\hat{G}\underline{D}} = B_{\hat{G}}$$

$$(28)\ B_{\tilde{G}\underline{M}} = B_{\tilde{G}\underline{D}} = B_{\tilde{G}}$$

$$(29)\ B_{\hat{X}\underline{M}} = B_{\hat{X}\underline{D}} = B_{\hat{X}}$$

$$(30)\ B_{\tilde{X}\underline{M}} = B_{\tilde{X}\underline{D}} = B_{\tilde{X}}$$

$$(31)\ B_{\hat{U}\underline{M}} = B_{\hat{U}\underline{D}} = B_{\hat{U}}$$

$$(32)\ B_{\tilde{U}\underline{M}} = B_{\tilde{U}\underline{D}} = B_{\tilde{U}}$$

If all these conditions hold, along with those specified in equations 18 through 24, the phenotypic variances for MZ and DZ twins should equal one another. They should also equal the phenotypic variances for ordinary siblings drawn from a similar population. In fact, these phenotypic variances do not differ significantly in our sample. But this is not *sufficient* proof that equations 18-24 and 27-32 hold. Equation 17 shows that the phenotypic variance is a complex function of the B's and s's. If there were reciprocal influence between twins, for example, the reduced-form value of $B_{\tilde{G}M}$ might exceed $B_{\tilde{G}D}$. But $s_{\tilde{X}M}$ might be less than $s_{\tilde{X}D}$, since a factor that ordinarily contributes to $s_{\tilde{X}}$, namely one's choice of friends outside the family, would be less important. Thus while equal variances are a necessary condition for accepting Assumptions I-III, they are not sufficient.

One further problem deserves attention. While the phenotypic variances for MZ pairs, DZ pairs, and siblings do not differ *significantly*, neither are they identical. If we assume that these differences are due to random sampling error, all the components of variance on the right side of equation 17 are equally likely to be affected. This implies that we can get an unbiased point estimate of heritability by standardizing variances for different sorts of twins and siblings to unity. If we use italics to distinguish these standardized variables and their regression coefficient, we can write our equation as:

$$(33)\ \underline{P_i} = \underline{a\hat{G}_i} + \underline{b\hat{X}_i} + \underline{c\hat{U}_i} + \underline{d\tilde{G}_i} + \underline{e\tilde{X}_i} + \underline{f\tilde{U}_i}$$

Taking variances we get:

$$(34)\ 1 = \underline{a}^2 + \underline{b}^2 + \underline{c}^2 + \underline{d}^2 + \underline{e}^2 + \underline{f}^2 + 2\underline{abr_{\hat{X}\hat{G}}}$$

Note that *a*, *b*, *c*, *d*, *e*, and *f* in equation 34 have no relation to a, b, c, d, e, and f in equations 1 to 9. The values of *a*, *b*, *c*, *d*, *e*, and *f* in equation 34 should be the same for MZ twins, DZ twins, and ordinary siblings.

If Assumptions I-III are correct, Figure 3 provides a full accounting of the sources of resemblance between twins reared together. Applying the basic theorem of path analysis, one can easily show that the phenotypic correlation between twins is:

(35) $r_{PP'} = a^2 + b^2 + c^2 + 2abm + nd^2 + pe^2 + qf^2$

where $m = r_{\hat{G}\hat{X}}$, $n = r_{\tilde{G}\tilde{G}'}$, $p = r_{\tilde{X}\tilde{X}'}$, and $q = r_{\tilde{U}\tilde{U}'}$. If Assumptions I-III are correct, a, b, c, d, e, f, and m have the same values for MZ and DZ pairs. But n, p, and q may differ for MZ and DZ pairs. If we use the subscripts M* and D* to designate parameters for fallibly diagnosed MZ and DZ pairs, we have:

(36M*) $r_{PP'M*} = a^2 + b^2 + c^2 + 2abm + n_{M*} d^2 + p_{M*}e^2 + q_{M*}f^2$

and

(36D*) $r_{PP'D*} = a^2 + b^2 + c^2 + 2abm + n_{D*}d^2 + p_{D*}e^2 + q_{D*}f^2$

Subtracting equation 36D* from equation 36M* then yields:

(37) $r_{PP'M*} - r_{PP'D*} = (n_{M*} - n_{D*})d^2 + (p_{M*} - p_{D*})e^2 + (q_{M*} - q_{D*}) f^2$

We want to solve equation 37 for d^2, the percentage of the phenotype variance due to genetic deviations from the family mean. We know that with perfect diagnoses of zygosity, $n_{M*} = 1$ and $n_{M*} = 0$. With errors in diagnosis one can show that:

(39) $n_{M*} - n_{D*} = 1 - E_{M*} - E_{D*}$

where E_{M*} and E_{D*} are the percentages of erroneously diagnosed MZ and DZ twins. It follows that:

(40) $d^2 = \dfrac{r_{PP'M*} - r_{PP'D*} - (p_{M*} - p_{D*})e^2 - (q_{DM*} - q_{DD*}) f^2}{1 - E_{M*} - E_{D*}}$

We know $r_{PP'M*}$ & $r_{PP'D*}$ and we can estimate E_{M*} and E_{D*}. In order to estimate d^2 we need two further assumptions:

IV MZ and DZ twins have equally correlated exogenous environments ($p_{M*} = p_{D*}$), and

V MZ and DZ twins have equally correlated interaction terms ($\underline{q}_{M*} = \underline{q}_{D*}$).

If these assumptions are correct, equation 40 reduces to

$$(41)\quad \underline{d}^2 = \frac{r_{\underline{PP'M*}} - r_{\underline{PP'D*}}}{1 - E_{M*} - E_{\underline{D}*}}$$

or with perfect diagnoses of zygosity,

$$(41a)\quad \underline{d}^2 = r_{\underline{PP'M}} - r_{\underline{PP'D}}$$

Assumption IV (Equally Correlated Exogenous Environments). This is by far the most controversial assumption in twin research. Environmentalists almost all argue that parents treat MZ twins more alike than DZ twins. The fact that parents treat MZ twins more alike than DZ twins is not, however, sufficient to prove that conventional heritability estimates are biased. First, parental treatment may have little impact on twin resemblance. Loehlin and Nichols (1976) show, for example, that while parents treated MZ twins in their sample more alike than DZ twins, the degree of similarity in treatment had no consistent or significant effect on the degree of resemblance between either MZ or DZ pairs on test scores or personality measures. Second, even if parental treatment does affect a given outcome, this need not bias our heritability estimates. Project Talent, for example, asked each respondent how much education he or she thought his or her parents expected her to get. MZ twins' responses to this question are more alike than DZ twins' responses. But parents' preferences regarding a child's educational attainment are not completely exogenous. They depend partly on the parents' perceptions of the child's own characteristics. If a child has great difficulty in school, for example, the parents are less likely to expect him to attend college than if the child learns easily. If a child is female rather than male, the parents are also less likely to expect him or her to attend college. This means that the child's genotype can affect the parents' expectations. It follows that parents will ordinarily have more nearly similar expectations for MZ twins than DZ twins, simply because MZ twins' genotypes are always the same whereas DZ twins' genotypes usually differ somewhat.

The question, then, is not whether the correlation between all environmental factors ($r_{EE'}$) is greater for MZ than DZ twins, but whether the correlation between exogenous environmental factors ($r_{XX'}$) is greater for MZ than for DZ twins. This poses a difficult empirical problem, for it is impossible to identify specific environmental measures that are clearly exogenous and yet vary within the same family. Many sociological measures, such as race, region, community size, parental income, father's occupation, and mother's education, are clearly exogenous. But these measures necessarily have the same value for twins reared together, regardless of zygosity. Other measures, such as how much the parents talk to the child, what games the parents encourage the child to play, or where they send the child to school, need not be the same for twins reared together. But a twin's genotype can affect all such measures, so we cannot be sure they are completely exogenous. Unless we have some way of knowing how much impact a child's genotype ordinarily has on a given environmental measure, we cannot say whether genotypic resemblance between MZ twins is sufficient to explain the fact that the overall environmental correlation is higher for MZ pairs than for DZ pairs.

The issue can perhaps be clarified by conducting a mental experiment. Suppose we could actually measure genotypes for both test scores (G_A) and education (G_B). If we did this, we would inevitably find a few DZ pairs with virtually identical predicted test scores. For these pairs $G_A = G'_A$, even though the specific genes involved are unlikely to be the same. But because the twins do not have the same genes, they are not likely to have identical genotypic values for outcomes other than the test scores. Unless genes affecting education were precisely the same as those affecting the test scores, DZ pairs with $G_A = G'_A$ would not usually have $G_B = G'_B$. If the pair were MZ, however, not only would $G_A = G'_A$, but $G_B = G'_B$. The same would hold for <u>all</u> genotypic outcomes. Now let us suppose we can also measure the environmental factors that affect test scores (E_A). We expect the environmental difference between two members of a pair ($E_A - E'_A$) to depend on the genetic difference with respect to test scores ($G_A - G'_A$). The question is whether it will <u>also</u> depend on the genetic difference with respect to education ($G_B - G'_B$) once we control $G_A - G'_A$. If it does, $r_{XA,\ XA'}$ is higher for MZ than for DZ pairs, since any environmental determinant of A that is not determined by G_A is by our definition exogenous, even if it depends on G_B. More generally, we can ask whether twins who resemble each other on a wide range of "irrelvant" genetic indices as well as on the relevant index end up in more similar environments than twins who resemble each other on the relevant index alone.

In order to illustrate the logic of this argument concretely, let us consider the effects of sex differences. Sex has a statistically insignificant effect on test scores in our sample, so $\bar{G}$ is virtually the same for males and females. It follows that the expected value of $r_{GG'}$ is the same for same-sex DZ pairs as for mixed-set DZ pairs. Yet it is easy to imagine that parents, teachers, and others treat same-sex DZ pairs more alike than mixed-set DZ pairs, especially if the same-sex pairs also happen to look quite a lot alike. This similarity in treatment cannot be endogenous since sex is not a component of test score genotype. Let us denote the observed phenotypic test score correlations for same-sex pairs as $r_{AA'\underline{D}*}$ and the observed phenotypic correlation for mixed-sex pairs as

$r_{AA'D''}$. Let us denote the phenotypic correlation for same-sex pairs after correcting for errors in diagnosing zygosity as $r_{AA'\underline{D}}$. For test scores, $r_{AA'\underline{D}} = 0.689$ while $r_{AA'\underline{D}''} = .416$. Despite the small sample size, this difference is just significant at the 0.05 level using a one-tailed test. Because the difference is only marginally significant, we also checked the value of $r_{AA'\underline{D}}$ for Schoenfeldt's larger sample of same-sex DZ twins. This sample included 98 pairs with incomplete followup data who had been excluded from our final sample. $r_{AA'\underline{D}}$ was 0.485 for the 98 same-sex pairs with incomplete followup data. This is only slightly larger than the correlation for mixed-sex pairs. The difference between Schoenfeldt's entire same-sex DZ sample ($r_{AA'\underline{D}} = 0.575$) and our mixed-sex sample ($r_{AA'\underline{D}''} = 0.416$) is, however, still just significant at the 0.05 level using a one-tailed test.

Unlike test scores, educational attainment varies significantly by sex. Since sex explains 1.6 percent of the variance in educational attainment, the expected correlation between same-sex pairs ($r_{BB'\underline{D}}$) is about (2)(0.016) = 0.032 higher than the expected correlation for mixed-sex pairs ($r_{BB'\underline{D}''}$). After correcting for errors in diagnosing zygosity, $r_{BB'\underline{D}} = 0.546$ and $r_{BB'\underline{D}'} = 0.467$ -- a difference of 0.079. The unexplained discrepancy between $r_{BB'\underline{D}}$ and $r_{BB'\underline{D}''}$ is therefore 0.079 - 0.032 = 0.047. This discrepancy is in the expected direction, but it is far short of being statistically significant.

Strictly speaking, one cannot draw conclusions about the true value of $r_{AA'D} - r_{AA'D''}$ from the observed value of $r_{BB'D} - r_{BB'D''}$ or vice versa. We suspect, however, that if we had sufficiently large samples we would probably find that $r_{AA'\underline{D}} - r_{AA'\underline{D}''}$ and $r_{BB'\underline{D}} - r_{BB'\underline{D}''}$ were quite similar. Since our best estimate of this difference is 0.159 for test scores (using Schoenfeldt's full sample of same-sex DZ twins) and 0.047 for educational attainment (after correcting for the direct effects of sex on education), and since there is no strong a priori basis for assuming that test scores are either more or less susceptible to such influences than educational attainment, it seems reasonable to assume that in general $r_{PP'D}$ exceeds $r_{PP'D''}$ by about (0.159 + 0.047)/2 = 0.10.

The implications of this tentative conclusion are a matter for subjective judgment. Environmentalists are likely to argue that if sex resemblance can inflate the correlation between twins by 0.10 above the value expected on the basis of genotypic resemblance and growing up together, other equally "irrelevant" forms of genetic resemblance could easily inflate the correlation between MZ twins by an additional 0.10 or 0.20. Since $r_{AA'M}$ and $r_{BB'M}$ only exceed $r_{AA'\underline{D}''}$ and $r_{BB'\underline{D}''}$ by about 0.20 in our samples, environmentalists are likely to conclude that heritability could easily be zero. Hereditarians, in contrast, are likely to argue that we have overestimated the difference between same-sex and mixed-sex twins, and that in any event the effects of sex resemblance are likely to be much greater than the effects of other sorts of "irrelevant" genetic resemblance.

The best way to resolve this issue is to measure other forms of genetic resemblance between DZ twins directly, and see whether they affect the phenotypic correlation between outcomes of interest. Carter-Saltzman and Scarr-Salapatek (1975) report, for example, that the IQ correlation between DZ twins whose parents believe they are identical does not differ significantly from the IQ correlation between DZ twins whose parents believe they are fraternal. Scarr (in correspondence) also reports that twins' test scores differ as much when they look very much alike as when they look quite different, once one controls true zygosity (determined from blood grouping). This sample is small, so the absence of a significant difference is hardly conclusive. Nonetheless, if we take Scarr's data at face value, we must conclude that the effects of sex resemblance on IQ are atypical and that most forms of physical resemblance do not affect $R_{XX'}$. If this holds for other "irrelevant" genetic factors (i.e. factors that do not affect G_A), and if it holds for educational attainment as well as IQ, we would have to conclude that the higher correlation between MZ twins' reports of their parents' educational expectations reflects the dependence of parental expectations on educational genotype. We do not have great confidence in this assumption. But it seems safer to overgeneralize Scarr's results than to rely entirely on guesswork. We will therefore assume that $r_{\tilde{X}\tilde{X}'\underline{M}} = r_{\tilde{X}\tilde{X}'\underline{D}}$ and that $\underline{p}_{\underline{M}*} = \underline{p}_{\underline{D}*}$.

Assumption V (MZ and DZ twins have equally correlated interaction terms). Let us begin by making the extreme assumption that pairs of twins always encounter the same environment, regardless of whether they are identical or fraternal. Thus $X_1 = X_1'$, $X_2 = X_2' \ldots X_n = X_n'$. Discrepancies between MZ twins' outcome measures would then be entirely due to measurement error, while discrepancies between DZ twins would be due to the combined effects of measurement error and genetic differences. MZ twins would then have identical values on both G and X. Not only that, but they would have identical values on the components of G and X, i.e., $G_1, G_2 \ldots G_n$ and $X_1, X_2 \ldots X_n$. This would ensure identical values on all the interaction terms involving $G_1 \ldots G_n$ and $X_1 \ldots X_n$. On the average DZ twins would have only half their genes in common. It follows that they would not have identical genes even if they happpened to have identical values on G. Thus if there were interactions between $G_1 \ldots G_n$ and $X_1 \ldots X_n$, DZ pairs would have identical values on only half these interaction terms. The correlation between $\tilde{U}$ and $\tilde{U}'$ would therefore be less than unity for DZ twins.

Now let us make the opposite assumption, namely that both MZ and DZ twins encounter random environments. Under these circumstances we would not expect to find any correlation between $\tilde{U}$ and $\tilde{U}'$ for either MZ or DZ twins. This suggests that $r_{\tilde{U}\tilde{U}'\underline{M}} = r_{\tilde{U}\tilde{U}'\underline{D}}$ if, and only if, $r_{\tilde{U}\tilde{U}'\underline{M}} = 0$

and $r_{\tilde{U}\tilde{U}'D} = 0$. This will happen if twins encounter random exogenous environments or if $s_{\tilde{U}M} = s_{\tilde{U}D} = 0$, i.e., if there are no interactions. Since neither assumption seems very plausible, one must assume that $r_{\tilde{U}\tilde{U}'M} > r_{\tilde{U}\tilde{U}'D}$, or in the notation of our path model, $q_M > q_D$.

In order to illustrate this phenomenon concretely, consider the combined effects of sex and birth order. Sex is completely genetic, while birth order is a completely exogenous environmental variable. Suppose that on the average neither sex nor birth order has any impact on some outcome. Suppose, however, that eldest male children perform better than average, while eldest females perform worse than average. Suppose this also holds for twins with no older siblings. Since birth order has no mean effect on twins' performance, $\tilde{X}$ is the same for first-born twins as for others. But the fact that same-sex twins share both their ordinal position and their sex genes would make their phenotypes more alike than those of mixed-sex twins. And since identical twins are always of the same sex, $r_{PP'M}$ would exceed $r_{PP'D+D''}$. This could be because $r_{\tilde{U}\tilde{U}'M}$ exceeded $r_{\tilde{U}\tilde{U}'D+D''}$.

Unfortunately, we have no way of knowing how common such phenomena are with respect to genes and environmental influences we cannot measure. Thus we cannot say how much larger $r_{\tilde{U}\tilde{U}'M}$ is likely to be than $r_{\tilde{U}\tilde{U}'D}$. It follows that we cannot solve equation 37 for d^2. The only way out of this difficulty is to assume that $r_{\tilde{U}\tilde{U}'M} = r_{\tilde{U}\tilde{U}'D}$ (i.e., $q_M = q_D$) and then recognize that this assumption inflates our estimate of heritability. Equation 41 then becomes:

$$d^2 \leq \frac{r_{PP'M*} - r_{PP'D*}}{1 - E_{M*} - E_{D*}} \tag{42}$$

What does this tell us? d is the unstandardized regression coefficient of $\tilde{G}$ when predicting P. Thus if we scale $\tilde{G}$ so that $d = 1$, $d = s_G / s_P$ and $s_G^2 = d^2 s_P^2$. If we know $s_{\tilde{G}}^2$ (i.e., the within-pair genetic variance) we can make a reasonable approximation of s_G^2 from theoretical considerations. Four such considerations are relevant.

(1) If parents mated randomly, if all genetic effects were additive, and if we scaled genotypes independently of the family mean, Mendelian inheritance would ensure that half the variance of genotypes fell within families and half fell between. $s_{\tilde{G}}^2 / s_G^2$ would then be 0.50.

(2) If there is positive assortative mating for a trait, as there is for test performance, education, and presumably career plans, the percentage of genotypic variance falling between families tends to rise.[16]

(3) If the effects of genes are not completely additive, i.e., if there is dominance or epistasis, less genetic variance will fall between families.[17]

The second and third considerations tend to cancel out. We therefore assume that about half the genotypic variance is between families.[18] This percentage could easily fall anywhere between 40 and 60 percent.

(4) If parents allocate more resources to their genetically disadvantaged children, as suggested earlier, and if they do this to a greater extent than society as a whole, a specified genetic difference between two children in the same family will have less ultimate impact than an identical difference between the means for two families. If this were the case, $\underline{d}$ would be reduced. Unfortunately, we have no evidence as to the likely magnitude of this bias. Indeed, we do not even know its direction with certainty, since in theory society could do more than families do to equalize their children. In the absence of evidence we will assume that this bias does not force $s^2_{\tilde{G}} / s^2_G$ below 0.40 or above 0.60.

These considerations suggest that if we use k to denote the presumed ratio of $s^2_{\tilde{G}}$ to s^2_G, we can estimate heritability using the formula:

$$(43)\quad \underline{h}^2 = \frac{s^2_G}{s^2_P} = \frac{s^2_G}{s^2_P} = \frac{s^2_{\tilde{G}}}{ks^2_P} = \frac{\underline{d}^2}{k} = \frac{r_{PP'\underline{M}*} - r_{PP'\underline{D}*}}{k(1 - E_{\underline{M}*} - E_{\underline{D}*})}$$

We expect k to fall between 0.40 and 0.60. With perfect diagnoses, this formula reduces to:

$$(43a)\quad \underline{h}^2 = \frac{r_{PP'\underline{M}} - r_{PP'\underline{D}}}{k}$$

Equation 43a is equivalent to Jensen's 1967 formula for heritability, namely:

$$(43b)\quad \underline{h}^2 = \frac{r_{PP'\underline{M}} - r_{PP'\underline{D}}}{r_{GG'\underline{M}} - r_{GG'\underline{D}}}$$

The equivalence can be demonstrated by recalling that with perfect diagnoses:

$$(44)\quad r_{GG'\underline{M}} = 1$$

and

$$(45)\quad r_{GG'\underline{D}} = \frac{s^2_{\hat{G}}}{s^2_G} = \frac{s^2_G - s^2_{\tilde{G}}}{s^2_G} = 1 - \frac{s^2_{\tilde{G}}}{s^2_G} = 1 - k$$

so

(46) $r_{GG'\underline{M}} - r_{GG'\underline{D}} = 1 - (1-k) = k$

The sampling variance of $\underline{h}^2$ in equation 43 depends on the sampling variances of $r_{PP'\underline{M}*} - r_{PP'\underline{D}*}$ and $1 - E_{\underline{M}*} - E_{\underline{D}*}$. Let us designate $r_{PP'\underline{M}*} - r_{PP'\underline{D}*}$ as Δr, $1 - E_{M*} - E_{D*}$ as D, and $1/(1 - E_{\underline{M}*} - E_{\underline{D}*})$ as E. Thus $E = 1/D$. Then:

(47) $k\underline{h}^2 = (\Delta r)(E)$.

Taking variances yields:

(48) $k^2 \text{ Var}(\underline{h}^2) = \text{Var}(\Delta rE)$

Using Goodman's (1960) formula for the variance of a product we get:

(49) $k^2\text{Var}(\underline{h}^2) = (\overline{\Delta r})^2 s_E^2 + \bar{E}^2 s_{\Delta r}^2 + s_{\Delta r}^2 s_E^2$. We do not know s_E^2, but $s_D^2 = s_{E\underline{M}*}^2 + s_{E\underline{D}*}^2$.

To express s_E in terms of s_D, we take the first derivative of D with respect to E. Since $D = E^{-1}$, $d(E)/d(D) = -E^{-2}$. This tells us that a one standard deviation change in D near the mean of the distribution of D will alter E by about $s_D/(\bar{D})^2$. This suggests that $s_E^2 \sim s_D^2/(\bar{D})^4$. Substituting, we get:

(50) $$k^2\text{Var}(\underline{h}^2) = (\overline{\Delta r})^2 s_D^2 + \frac{s_{\Delta r}^2}{\bar{D}^2} + \frac{s_{\Delta r}^2 s_D^2}{\bar{D}^4}$$

where $s_D^2 = s_{E\underline{M}*}^2 + s_{E\underline{D}*}^2$ and our best estimate of $\bar{D}$ is $1 - E\underline{M}* - E_{\underline{D}*}$.

The best estimate of $\overline{\Delta r}$ is $r_{PP'\underline{M}*} - r_{PP'\underline{D}*}$. The sampling variance of Δr depends on the sampling errors of $r_{PP'\underline{M}*}$ and $r_{PP'\underline{D}*}$. These are not normally distributed. But Fisher's z transformation of r, i.e. $z = 0.5 \ln \left[(1+r)/(1-r)\right]$, is almost normally distributed with a variance of $1/(N-3)$, where N is the number of pairs on which the original correlation is based. Since $z_{PP'\underline{M}*}$ and $z_{PP'\underline{D}*}$) are normally distributed, Δz (i.e. $z_{PP'\underline{M}*} - z_{PP'\underline{D}*}$) should also be normally distributed. Furthermore, since $z_{PP'\underline{M}*}$ and $z_{PP'\underline{D}*}$ are independent, $s_{\Delta z}^2 = 1/(N_{\underline{M}*} - 3) + 1/(N_{\underline{D}*} - 3)$. We can therefore calculate $\Delta z/s_{\Delta z} = t_\Delta$. To get $s_{\Delta r}$, we assume that $\Delta r/s_{\Delta r} = \Delta z/s_{\Delta z} = t_\Delta$.

It follows that $s^2_{\Delta r} = (\Delta r)^2/t^2_\Delta$.

Substituting the foregoing values in equation 50 gives us:

$$(51)\ k^2 \mathrm{Var}(\underline{h}^2) = (\Delta r)^2 (s^2_{E\underline{M}*} + s^2_{E\underline{D}*}) + \frac{(\Delta r)^2}{t^2_\Delta (1 - E_{\underline{M}*} - E_{\underline{D}*})^2}$$

$$+ \frac{(\Delta r)^2 (s^2_{E\underline{M}*} + s^2_{E\underline{D}*})}{t^2_\Delta\ (1 - E_{\underline{M}*} - E_{\underline{D}*})^4}$$

$$= \frac{(t^2_\Delta + 1)(\Delta r)^2 (s^2_{E\underline{M}*} + s^2_{E\underline{D}*}) + (\Delta r)^2 (1 - E_{\underline{M}*} - E_{\underline{D}*})^2}{t^2_\Delta (1 - E_{\underline{M}*} - E_{\underline{D}*})^4}$$

Taking square roots then yields:

$$(52)\ ks_{\underline{h}2} = \frac{\Delta r \left[(t^2_\Delta + 1)(s^2_{E\underline{M}*} + s^2_{E\underline{D}*}) + (1 - E_{\underline{M}*} - E_{\underline{D}*})^2\right]^{1/2}}{t_\Delta (1 - E_{\underline{M}*} - E_{\underline{D}*})^2}$$

But $\Delta r/(k)(1 - E_{\underline{M}*} - E_{\underline{D}*}) = \underline{h}^2$, so:

$$(53)\ s_{\underline{h}2} = \frac{\underline{h}^2}{t_\Delta} \left(\frac{(t_\Delta + 1)(s^2_{E\underline{M}*} + s^2_{E\underline{D}*})}{(1 - E_{\underline{M}*} - E_{\underline{D}*})^2} + 1 \right)^{1/2}$$

If blood samples are available and all twins are diagnosed accurately, $E_{M*} = E_{D*} = \mathrm{Var}(E_{M*} + E_{D*}) = 0$. Equation 50 then reduces to

$$(54)\ s_{\underline{h}2} = \frac{\underline{h}^2}{t_\Delta}$$

Schoenfeldt estimated $E_{\underline{M}*}$ and $E_{\underline{D}*}$ from a sample collected by Nichols rather than Talent in which $N_{M*} = 83$ and $N_{D*} = 41$. We will use his results to estimate the probable values of E_{M*} and E_{D*} in our Talent samples. $s^2_{E\underline{M}*} + s^2_{E\underline{D}*}$ works out to be 0.074^2 for Schoenfeldt's original

Talent sample, 0.083^2 for our sample with followup data, and 0.094^2 for our subsample of pairs who both report career plans.

IV. Heritability Estimates for Single Traits

Table 4 gives our "maximum estimates" of $\bar{h}^2$, along with the approximate standard errors of these maximums. We present estimates of $\bar{h}^2$ for three possible values of k, i.e. k = 0.040, 0.50, and 0.60. The actual value of k depends on the degree of dominance and assortative mating for each trait, and on whether families make more effort than society as a whole to compensate individuals for genetic disadvantages. If assortative mating is strong, if dominance is unimportant, and if families engage in effective compensation, k will be high. If dominant genes are important, if assortative mating on genotype is weak, and if families compensate genetic disadvantages no more than society as a whole, k will be low. Because $r_{XX'}$ appears to be lower for mixed-sex than for same-sex pairs, we have estimated $\bar{h}^2$ by comparing same-sex MZ pairs to same-sex DZ pairs, ignoring mixed-sex pairs.

Test Scores. Table 4 presents two sets of heritability estimates for Talent test scores. The first series covers all same-sex pairs of known zygosity in Schoenfeldt's original sample. The second series covers our subsample of Talent twins with 5-year followup data. As Table 2 indicated, the correlation between same-sex DZ twins' test scores is appreciably higher in our sample with followup data than in Schoenfeldt's original sample. Our sample therefore implies lower values of $\bar{h}^2$ than Schoenfeldt's sample. Using what we regard as the most likely value of k, namely 0.50, the estimate of $\bar{h}^2$ is 0.387 ± 0.144 for our subsample, compared to 0.581 ± 0.125 for Schoenfeldt's initial sample. Since Schoenfeldt's sample is larger than ours and is less subject to selection bias, the 0.581 figure is clearly preferable. The estimates based on Schoenfeldt's sample are consistent with those obtained by most other investigators.[19]

Education. Table 4 also estimates the heritability of educational attainment in our Talent sample and in Taubman's much larger sample of white male twins born 1917-27.[20] The two sets of heritability estimates are remarkably similar, despite the fact that they come from different target populations.

The fact that half the Talent sample is female has two contradictory effects on $\bar{h}^2$. First, Table 2 shows that if k = 0.50, $\bar{h}^2$ = 0.540 for males and 0.450 for females. This difference could be due to chance, but it suggests that at least in this sample including females should lower $\bar{h}^2$. This bias is offset, however, by the fact that the variance for females is restricted and the fact that sex itself affects education. As a result, the heritability estimate for males is almost identical to that for the full sample. In theory, the Talent sample also differs from Taubman's by including non-whites, but in fact only two percent of our final sample was non-white.

The similarity between our results for males and Taubman's therefore suggests that there was very little change in the contribution of genetic

Table 4:

Maximum Estimates of h^2 for Test Scores, Education, and Career Plans (With Standard Errors in Parentheses), Using Three Alternative Assumptions About the Combined Effects of Dominance, Assortative Mating, and Familial Resource Allocation on k.

	k		
	0.60	0.50	0.40
Test Scores			
Schoenfeldt Talent Sample (Same-Sex Pairs)	.484 (.104)	.581 (.125)	.726 (.156)
Jencks-Brown Talent Sample (Same-Sex Pairs)	.323 (.120)	.387 (.144)	.484 (.180)
Education			
Taubman NAS-NRC Sample (White Males Born 1917-27)	.407 (.046)	.489 (.055)	.611 (.069)
Jencks-Brown Talent Sample (Same-Sex Pairs)	.432 (.173)	.518 (.207)	.648 (.259)
Career Plans at 23			
Jencks-Brown Talent Sample (Same-Sex Pairs)	1.039 (.399)	1.247 (.479)	1.559 (.599)

variation to educational inequality between the time Taubman's cohort finished school (roughly 1931 to 1950) and the time the Talent cohort finished (roughly 1960 to 1968). The mean level of education rose sharply during this interval, and the variance fell. But the relative importance of genetic and non-genetic factors apparently remained unchanged. This conclusion is consistent with Blau and Duncan's (1967) finding that the correlation between educational attainment and parental status was stable for cohorts born between 1897 and 1936. Hauser and Featherman (1975) found that this stability persisted for the cohort born between 1937 and 1946.

Career Plans. Our "best" point estimate of h^2 for career plans is 1.247. This is embarrassing, because the population value of h^2 cannot exceed unity. Given the large sampling errors of our estimates, some may not take this difficulty too seriously. The illegitimate point estimate does, however, point to one of the peculiar difficulties of heritability estimation. Heritability estimates require us to combine data from two independent samples of MZ and DZ twins. Yet they need not yield point estimates that fall between those for the two samples. In order to see that this is so, look back at Figure 3. The correlation between any given pair is

$$(55)\quad r_{PP'} = a^2 + b^2 + c^2 + 2abm + nd^2 + pe^2 + qf^2$$

Since m, n, p, and q are virtually certain to be positive and since all the squared terms must be positive, it is safe to assume that:

$$(56)\quad r_{PP'} > a^2 + nd^2$$

For correctly diagnosed MZ pairs, $n = 1$. For correctly diagnosed DZ pairs, $n = 0$. For all pairs, $a^2 + d^2 = h^2$. Furthermore, if $k = 0.50$, $a^2 = 0.50h^2$. It follows that:

$$(57)\quad h^2 < r_{PP'M}$$

and

$$(58)\quad h^2 < 2r_{PP'D}$$

After correcting for misdiagnosis, $r_{PP'M} = 0.644$ and $r_{PP'D} = 0.020$ for our sample. The implied maxima for h^2 are thus 0.644 for MZ pairs and 0.04 for DZ pairs. Nonetheless, we end up with a point estimate of $h^2 = 1.247$.

The reason is that no legitimate value of h^2 will allow us to assume the same between-family environmental variance (b^2) for both MZ and DZ pairs and then reproduce the observed phenotypic correlations between MZ and DZ twins' career plans. The only way to reproduce these correlations is to assume an unreasonably large value of h^2 (e.g. 1.247) and then assume a negative value for b^2 (e.g. -0.603). This is plainly unreasonable. The most realistic alternative is to assume that random sampling error has given us a DZ sample with unusually low values of $\underline{h}^2$, $\underline{b}^2$, or both. This assumption is supported by the fact that the phenotypic correlations for mixed-sex DZ twins and for siblings are higher than the observed phenotypic correlation for same-sex DZ twins, despite the fact that the same-sex twin sample presumably includes a few misdiagnosed MZ pairs. But even if we assume $r_{PP'D}$ = 0.25 (the observed r when we combine twins and siblings), we get heritability in excess of 0.80. This still exceeds the maxima implied by equations 57 and 58, which are 0.65 and 0.50 respectively. This suggests that sampling error has also given us an MZ sample with an unusually high value for $\underline{h}^2$ or $\underline{b}^2$. Heritability estimates that rely on comparing samples drawn from opposite extremes of their respective sampling distributions are unlikely to prove enlightening. For this reason we will not discuss the results for Career Plans any further.

V. Are the Genes that Affect Education the Same as Those that Affect Test Performance?

Until recently, virtually all investigators assumed that if genes affected life chances at all, they did so primarily by affecting "intelligence."[21] Yet the one set of genes we can actually identify in this survey, namely those that determine sex, does not operate in this fashion. Sex has virtually no impact on test performance in the Talent sample, yet it clearly affects educational attainment. This section tries to estimate the extent to which other genes operate in this same fashion, affecting educational attainment independently of test performance.

We begin with a simple calculation. Suppose that the broad heritabilities in Table 4 are unbiased and that k = 0.50. Table 4 then implies that genes explain 38.7 percent of the variance in test scores in our sample. Table 2 shows that test scores explain 30.4 percent of the variance in educational attainment. It follows that if genes have no direct effect on educational attainment, they cannot explain more than (0.387) (0.304) = 11.8 percent of the variance in educational attainment in our sample. Yet Table 4 suggests that genes actually explain 51.8 percent of the variance in educational attainment in our sample. If our assumptions are at all reasonable, then, genes must have a substantial direct effect on educational attainment in our sample independent of test scores. (This conclusion might not hold if we assumed that twins affected one another's educational attainment but did not appreciably affect one another's test scores. Then the true education heritability would be much lower than in Table 4, but the test score heritability would not.)

If genes have appreciable effects on educational attainment among respondents with the same test scores, we can ask whether the genes that affect educational attainment are the same ones that affect test performance. If exactly the same genes are involved, the correlation will be 1.00. If completely different genes are involved correlations between G_A and G_B will be zero.

We could estimate $r_{GA,\ GB}$ using the same assumptions and notation as in section III, but the estimation procedure will be clearer if we simplify the earlier notation. Let:

$$(59)\quad G_{\underline{i}} = \hat{G}_i + \tilde{G}_i$$

and

$$(60)\quad \underline{U}_i = \hat{X}_{\underline{i}} + \tilde{X}_{\underline{i}} + \hat{U}_i + \tilde{U}_{\underline{i}}$$

It follows that:

$$(61)\quad P_{\underline{i}} = G_{\underline{i}} + U_{\underline{i}}$$

and

$$(61')\quad P'_{\underline{i}} = G'_{\underline{i}} + U'_{\underline{i}}$$

Taking variances, we have:

$$(62)\quad \mathrm{Var}P = \mathrm{Var}G + \mathrm{Var}\underline{U} + 2\mathrm{Cov}G\underline{U}$$

If we decompose $\mathrm{Cov}G\underline{U}$ into the covariances of the components of G with the components of $\underline{U}$, these components are all zero except $\mathrm{Cov}\hat{G}\hat{X}$, which is likely to be positive. Thus:

$$(63)\quad \mathrm{Var}P = \mathrm{Var}G + \mathrm{Var}\underline{U} + 2\mathrm{Cov}\hat{G}\hat{X}$$

Multiplying equation 61 by 61', summing over all observations, and dividing by the number of observations we get:

$$(64)\quad \mathrm{Cov}PP' = \mathrm{Cov}GG' + \mathrm{Cov}\underline{UU'} + 2\mathrm{Cov}G\underline{U'}$$

Writing variants of this equation for fallibly diagnosed MZ and DZ twins and subtracting the latter from the former we have:

$$(65)\quad \mathrm{Cov}PP'_{\underline{M}*} - \mathrm{Cov}PP'_{\underline{D}*} = \mathrm{Cov}GG'_{\underline{M}*} - \mathrm{Cov}GG'_{\underline{D}*} + \mathrm{Cov}\underline{UU'}_{\underline{M}*} - \mathrm{Cov}\underline{UU'}_{\underline{D}*} + 2\mathrm{Cov}G\underline{U'}_{\underline{M}*} - 2\mathrm{Cov}G\underline{U'}_{\underline{D}*}$$

If the substantive assumptions discussed in section III are all correct,

$\mathrm{Cov}\underline{UU'}_{\underline{M}*} = \mathrm{Cov}\underline{UU'}_{\underline{D}*}$ and $\mathrm{Cov}\underline{GU'}_{\underline{M}*} = \mathrm{Cov}GU'_{\underline{D}*}$, giving us:

$$(66)\quad \mathrm{Cov}PP'_{\underline{M}*} - \mathrm{Cov}PP'_{\underline{D}*} = \mathrm{Cov}GG'_{\underline{M}} - \mathrm{Cov}GG'_{\underline{D}*}$$

or

$$(67)\ (r_{PP'M*} - r_{PP'D*})\ VarP = (r_{GG'M*} - r_{GG'D*})\ VarG$$

We can rearrange this as:

$$(68)\ h^2 = \frac{VarG}{VarP} = \frac{r_{PP'M*} - r_{PP'D*}}{r_{GG'M*} - r_{GG'D*}}$$

This is Jensen's (1967) formula. It can also be written as:

$$(69)\ h^2 = \frac{r_{PP'M*} - r_{PP'D*}}{(1 - r_{GG'D})\ (1 - E_{M*} - E_{D*})}$$

where E_{M*} and E_{D*} are the percentages of apparently MZ and DZ twins that are actually DZ and MZ respectively. $1 - r_{GG'D}$ presumably falls between 0.40 and 0.60. Equation 69 is identical to equation 43 (our earlier estimate of h^2) except that it substitutes $1 - r_{GG'D}$ for k. This substitution involves no substantive change, since $1 - r_{GG'D} = 1 - s^2_{\hat{G}}/s^2_G = s^2_{\tilde{G}}/s^2_G = k$.

In order to extend this analysis to two outcomes, we denote the phenotypic values as A and B, the genotypic values as G_A and G_B, and the combined effects of exogenous environment and genotype-environment interactions as U_A and U_B. This allows us to write:

$$(70)\ A_i = G_{Ai} + U_{Ai}$$

and

$$(71)\ B'_i = G'_{Bi} + U'_{Bi}$$

Multiplying equation 70 by equation 71, summing over all pairs, and dividing by the number of pairs, we get:

$$(72)\ CovAB' = CovG_A G'_B + CovG_A U'_B + GovU_A G'_B + CovU_A U'_B$$

If we write variants of equation 72 for MZ and DZ twins and subtract the latter from the former, we have:

$$(73)\ Cov_{M*}AB' - Cov_{D*}AB' = Cov_{M*}G_A G_B - Cov_{D*}G_A G_B + Cov_{M*}G_A U'_B -$$

$$Cov_{D*}G_A U'_B + Cov_{M*}U_A G'_B - Cov_{D*}U_A G'_B + Cov_{M*}U_A U_B - Cov_{D*}U_A U'_B$$

If we make the same substantive assumptions as before, we can show that $Cov_{\underline{M}*}G_A\underline{U}'_B = Cov_{\underline{D}*}G_A\underline{U}'_B$, that $Cov_{\underline{M}*}\underline{U}_AG'_B = Cov_{\underline{D}*}\underline{U}_AG'_B$, and that $Cov_{\underline{M}*}\underline{U}_AU'_B = Cov_{\underline{D}*}\underline{U}_A\underline{U}'_B$. Equation 73 then reduces to:

(74) $Cov_{\underline{M}*}AB' - Cov_{\underline{D}*}AB' = Cov_{\underline{M}*}G_AG'_B - Cov_{\underline{D}*}G_AG'_B$

or

(75) $(r_{AB'\underline{M}*} - r_{AB'\underline{D}*})s_As_B = (r_{GA,\,GB'\underline{M}*} - r_{GA,\,GB'\underline{D}*})s_{GA}s_{GB}$

But $s^2_{GA}/s^2_A = \underline{h}^2_A$ and $s^2_{GB}/S^2_B = \underline{h}^2_B$ so:

(76) $r_{AB'\underline{M}*} - r_{AB'\underline{D}*} = (r_{GA,\,GB'\underline{M}*} - r_{GA,\,GB'\underline{D}*})\underline{h}_A\underline{h}_B$

In order to put equation 76 to use, we must express $r_{GA,GB'}$, in terms of $r_{GA,GB}$, $r_{GAGA''}$ and $r_{GB,GB''}$. We first write:

(77) $G_{\underline{Ai}} = \hat{G}_{\underline{Ai}} + \tilde{G}_{\underline{Ai}}$

and

(77') $G'_{\underline{Bi}} = \hat{G}_{\underline{Bi}} + \tilde{G}'_{\underline{Bi}}$

Multiplying equation 77 by 77', summing over all pairs, dividing by N, and dropping the zero terms we get:

(78) $CovG_AG'_B = Cov\hat{G}_A\hat{G}_B + Cov\tilde{G}_A\tilde{G}'_B$

Rearranging, this yields:

(79) $$r_{GA,GB'} = \frac{r_{\hat{G}A,\hat{G}B}\,s_{\hat{G}A}\,s_{\hat{G}B}}{s_{GA}\,s_{GB}} + \frac{r_{\tilde{G}A,\tilde{G}B}\,s_{\tilde{G}A}\,s_{\tilde{G}B}}{s_{GA}\,s_{GB}}$$

We now assume that $s^2_{\hat{G}A}/s^2_{GA} \sim s^2_{\hat{G}B}/s^2_{GB} \sim 1 - k$, and $s^2_{\tilde{G}A}/s^2_{GA} \sim s^2_{\tilde{G}B}/s^2_{\tilde{G}B} \sim k$. This is equivalent to assuming that the effects of dominance, assortative mating and environmental compensation for genetic disadvantages are similar for the two outcomes. Then:

(80) $r_{GA,GB'} \sim r_{\hat{G}A,\hat{G}B}(1 - k) + r_{\tilde{G}A,\tilde{G}'B}\,k$

Since the twins inherit random genes from their parents, $r_{\hat{G}A,\hat{G}B} = r_{GA,GB}$. For correctly diagnosed MZ pairs, moreover,

$r_{\tilde{G}A,\tilde{G}B'} = r_{\tilde{G}A,\tilde{G}B} = r_{GA,GB}$. For correctly diagnosed DZ pairs, $r_{\tilde{G}A,\tilde{G}B'} = 0$. With errors in diagnosis, $r_{\tilde{G}A,\tilde{G}B'\underline{M}*} = r_{GA,GB}(1-E_{\underline{M}*})$ and $r_{\tilde{G}A,\tilde{G}B'\underline{D}*} = E_{\underline{D}*}r_{GA,GB}$. Substituting into equation 80 therefore yields:

(80M) $$r_{GA,GB'\underline{M}*} \sim r_{GA,GB}(1 - k) + r_{GA,GB}k(1 - E_{\underline{M}*})$$

and

(80D) $$r_{GA,GB'\underline{D}*} \sim r_{GA,GB}(1 - k) + r_{GA,GB}kE_{\underline{D}*}$$

Subtracting equation 80D from 80M we have:

(81) $$r_{GA,GB'\underline{M}*} - r_{GA,GB'\underline{D}*} \sim r_{GA,GB}k(1 - E_{\underline{M}*} - E_{\underline{D}*})$$

Substituting into equation 76 gives us

(82) $$r_{AB'\underline{M}*} - r_{AB'\underline{D}*} \sim r_{GA,GB}k(1 - E_{\underline{M}*} - E_{\underline{D}*})\,\underline{h}_A\underline{h}_B$$

But equation 43 showed that $\underline{h}_A^2 = (r_{AA'\underline{M}*} - r_{AA'\underline{D}*})/k(1 - E_{\underline{M}*} - E_{\underline{D}*})$ and $\underline{h}_B^2 = (r_{BB'\underline{M}*} - r_{BB'\underline{D}*})/k(1 - E_{\underline{M}*} - E_{\underline{D}*})$, so equation 82 reduces to:

(83) $$r_{AB'\underline{M}*} - r_{AB'\underline{D}*} \sim r_{GA,GB}(r_{AA'\underline{M}*} - r_{AA'\underline{D}*})^{\frac{1}{2}}(r_{BB'\underline{M}*} - r_{BB'\underline{D}*})^{\frac{1}{2}}.$$

Rearranging,

(84) $$r_{GA,GB} \sim \frac{r_{AB'\underline{M}*} - r_{AB'\underline{D}*}}{(r_{AA'\underline{M}*} - r_{AA'\underline{D}*})^{1/2}(r_{BB'\underline{M}*} - r_{BB'\underline{D}*})^{1/2}}$$

For test scores and education Table 2 implies:

$$r_{GA,GB} = \frac{0.586 - 0.544}{(0.864 - 0.722)^{1/2}(0.780 - 0.590)^{1/2}} = 0.268$$

We could calculate the sampling error of this estimate using maximum likelihood procedures, but we have not done this. The sampling error appears to be very large. (The implied value of $r_{GA,\ GB}$ for males in this sample, is 0.283, for example, while the implied value for females is 0.046.) We could also calculate the likely bias in this estimate if one or more of the assumptions stated in Section III were wrong. We have not done this either, but the bias could be substantial.

Our final question is the extent to which genes account for variations in educational attainment among individuals with similar test scores. This is not the same as our previous question, namely whether the same genes affect both outcomes. One could, after all, imagine a situation in which, say, the genes that determine skin color were the only genes that affected either test scores or educational attainment. This would make $r_{GA,GB} = 1$. But this would not tell us whether skin color affected educational attainment with test scores controlled.

In order to estimate genes' effects on education with test scores controlled, we write two new structural equations, i.e.

$$(85)\ A = \underline{a}G'_A + \underline{b}\hat{U}_A + u_A$$

and

$$(85')\ A' = \underline{a}G'_A + \underline{b}\hat{U}_A + u'_A$$

where A represents test performance, G represents those genes that affect test performance, $\hat{U}$ represents the family's expected deviation from its genotypic mean, and u_A represents an error term. $\underline{a}$ and $\underline{b}$ are standardized regression coefficients. Note that $\underline{a} = \underline{h}_A$ in our previous models. We now write:

$$(88)\ B = \underline{c}A + \underline{d}\underline{G}_B + \underline{e}\hat{\underline{U}}_B + u_B$$

and

$$(88')\ B' = \underline{c}A' + \underline{d}\underline{G}'_B + \underline{e}\hat{\underline{U}}_B + u'_B$$

where $\underline{G}_B$ represents the genes that affect education once we control ability, $\underline{U}_B$ represents the family's expected deviation from the value predicted on the basis of $\hat{A}$ and $\underline{G}_B$, and u_B is an error term. Note that $\underline{G}_B$ is not the same as G_B since it does not include those aspects of education genotype that operate through test scores. Likewise, $\hat{\underline{U}}_B$ is not the same as $\hat{U}_B$. $\underline{c}$, $\underline{d}$, and $\underline{e}$ are again standardized coefficients, but $\underline{d}$ is not equal to $\underline{h}_B$ in our previous models. Figure 4 presents this model visually.

Using the basic algorithm of path analysis, one can show that our four "structural" equations imply the following relationships:

$$(89)\ r_{AB} = \underline{c} + \underline{adr}_{GA,GB} + \underline{ber}_{\hat{U}A,\hat{U}B} + \underline{bdr}_{\hat{U}A,GB} + \underline{aer}_{GA,\hat{U}B}$$

$$(90)\ r_{AA'} = \underline{a}^2 r_{GA,GA'} + \underline{b}^2 + 2\underline{abr}_{GA,\hat{U}A}$$

$$(91)\ r_{AB'} = \underline{c}\, r_{AA'} + \underline{adr}_{GA,\underline{G}B'} + \underline{ber}_{\hat{U}A,\hat{U}B} + \underline{bdr}_{\hat{U}A,GB} + \underline{aer}_{GA,\hat{U}B}$$

$$(92)\ r_{BB'} = \underline{c}r_{AB'} + \underline{d}^2 r_{\underline{G}B,\underline{G}B'} + \underline{e}^2 + 2\underline{de}r_{\underline{G}B,\hat{\underline{U}}B} +$$

$$\underline{acd}r_{\underline{G}A',\underline{G}B} + \underline{bcd}r_{\hat{U}A,GB} + \underline{ae}r_{\hat{\underline{U}}B,GA'} + \underline{bce}r_{\hat{U}A,\hat{\underline{U}}B}$$

Before using these equations to estimate the parameter that interests us ($r_{GA,GB}$), three difficulties deserve comment.

(1) Equation 89 implies that the correlation between test scores and educational attainment is the same for MZ and DZ pairs. Yet the observed correlations are 0.621 for MZ pairs and 0.565 for same-sex DZ pairs. This difference could easily be due to sampling error, but it must still be dealt with in some way. We will return to this problem below.

(2) Equation 91 implies that $r_{AB'}$ is larger for MZ twins than for DZ twins. The observed values are 0.586 for MZ pairs and 0.544 for DZ pairs. This difference is in the expected direction, but it is smaller than the difference between r_{ABM*} and r_{ABD*}, which is supposedly due to chance alone. Sampling error will, of course, occasionally produce point estimates that follow this inconvenient pattern. But erecting an elaborate structure of inference on two such samples is unlikely to yield reliable conclusions about the real world.

(3) $\underline{c}$ is the coefficient of test scores when predicting educational attainment and controlling both genotype and common environment. One way to estimate $\underline{c}$ is to look at differences between MZ twins reared together. $\underline{c}$ is the expected difference in educational attainment between MZ twins whose test scores differ by one standard deviation. We can estimate this within-pair regression coefficient (b_Δ) from the general formula:

$$(93)\ \underline{b}_\Delta = \frac{r_{AB} - r_{AB'}}{1 - r_{AA'}}$$

Table 2 indicates that with errors in diagnosis $\underline{b}_\Delta$ is 0.257 for apparently MZ pairs and 0.076 for DZ pairs. The fact that $\underline{b}_\Delta$ is larger for MZ pairs than for DZ pairs is somewhat surprising. The within-pair regression for MZ pairs in effect controls all environmental influences common to both twins plus all genetic influences. The within-pair regression for DZ pairs controls all common environmental influences but only half the relevant genetic influences. One would expect the correlations between education genotypes and test scores to be positive even after controlling common environmental influences. One would therefore expect the within-pair coefficient for DZ pairs to be larger than the within-pair coefficient for MZ twins. Since it is not,either (a) education genotype is negatively correlated with test scores, after controlling exogenous environmental influences common to both twins, or (b) the within-pair DZ coeffieient is low simply because of sampling error. In order to assess the relative likelihood of these two explanations, we used equation 93 to estimate the standardized within-pair coefficients for mixed-sex DZ pairs, same-sex siblings, and mixed-sex siblings. These coefficients were all higher than the coefficient for MZ pairs, just as theory implies they should be. We therefore concluded that the same-sex DZ sample was aberrant. It follows that inferences based on comparisons between the MZ sample and the same-sex DZ sample are likely to be misleading.

In light of these difficulties we will not try to draw any substantive conclusions about the effects of genotype on educational attainment among respondents with identical test scores. We will, however, present the equations necessary for estimating such effects from more representative samples. We begin by estimating c, the within-pair coefficient for correctly diagnosed MZ pairs. This can be derived from equation 13 as

$$(94)\quad c = b_{\Delta M^*} + \frac{E_{M^*}(b_{\Delta M^*} - b_{\Delta D^*})}{1 - E_{M^*} - E_{D^*}}$$

Next, we write variants of equations 90–93 for MZ and DZ twins, and subtract the DZ equations from their MZ analogs. This gives us:

$$(95)\quad r_{AA'M^*} - r_{AA'D^*} = a^2(r_{GA,GA'M^*} - r_{GA,GA'D^*})$$

$$(96)\quad r_{AB'M^*} - r_{AB'D^*} = ad(r_{GA,GB'M^*} - r_{GA,GB'D^*}) + c(r_{AA'M^*} - r_{AA'D^*})$$

$$(97)\quad r_{BB'M^*} - r_{BB'D^*} = c(r_{AB'M^*} - r_{AB'D^*}) + d^2(r_{GB,GB'M^*} - r_{GB,GB'D^*}) + acd(r_{GA',GB'M^*} - r_{GA,GB'D^*})$$

Finally, we express $r_{GA,GA'M^*} - r_{GA,GA'D^*}$, $r_{GA,GB'M^*}$ $r_{GA,GB'M^*}$, and $r_{GB,GB'M^*} - r_{GB,GB'D^*}$ in terms of $r_{GA,GB}$. To do this we set $r_{GG'M^*} - r_{GG'D^*} = k$. If k is approximately the same for A and B, the same will hold for k. Using the logic that led to equation 81 we can then show that:

$$(98)\quad r_{GA,GB'M^*} - r_{GA,GB'D^*} = k r_{GA,GB}$$

Equations 95 – 97 then become:

$$(99)\quad a = \left(\frac{r_{AA'M^*} - r_{AA'D^*}}{k}\right)^{1/2}$$

$$(100)\ \underline{d}r_{GA,\underline{G}B} = \frac{r_{AB'\underline{M}*} - r_{AB'\underline{D}*} - \underline{c}(r_{AA'\underline{M}*} - r_{AA'\underline{D}*})}{\underline{ak}}$$

$$(101)\ \underline{kd}^2 + \underline{acd}r_{GA,\underline{G}B}\underline{k} + \underline{c}(r_{AB'\underline{M}*} - r_{AB'\underline{D}*}) - (r_{BB'\underline{M}*} - r_{BB'\underline{D}*}) = 0.$$

One can solve equation 94 for $\underline{c}$, then solve equation 99 for $\underline{a}$, then solve equation 100 for $\underline{d}r_{GA,GB}$, and then substitute this value of $\underline{d}r_{GA,\underline{G}B}$ into equation 101 and solve for $\underline{d}^2$. Note that this routine does not require maximum likelihood methods. Anyone with a small calculator can get solutions in ten minutes. We should note, however, that this routine for estimating the relevant parameters does not fully utilize r_{ABD*} (except in correcting $\underline{b}_{\Delta \underline{M}*}$ for errors in diagnosing zygosity). Instead, we implicitly assume that $r_{A\underline{B}D*} = r_{AB\underline{M}*}$. Had we not made this implicit assumption we would have had an overidentified system and would have needed maximum-likelihood procedures to obtain an efficient solution.

One can, of course, easily apply both equation 84 and equations 94, 99, 100, and 101 to Taubman et al's twin data.

VI. Conclusions

(1) A high correlation between genotype and phenotype need not imply a low correlation between environmental influences and phenotype. The estimated value of $\underline{h}^2$ sets no limit on the percentage of variance traceable to environmental inequality unless one can set some limit on the likely correlation between genotype and environment. Since genotypes can and do affect the environments human beings choose or have chosen for them, the correlation between genotype and environment broadly conceived may be as high as 1.00. Efforts at partitioning the phenotypic variance of human traits into hereditary and environmental components should therefore be abandoned. The most we can do is partition the variance into hereditary and non-hereditary components.

(2) Those who want to estimate the contribution of environmental inequality to phenotypic inequality must begin by offering some meaningful definition of what they mean by "environment". One possible approach is to define "environment" as including only those features of the environment that one has measured. If one fails to measure certain crucial environmental influences, one will then underestimate the overall impact of environment. But at least this possibility will be obvious to all concerned. The other possible approach to estimating the effects of environmental inequality is to look only at those environmental influences shared by children reared together. One can estimate the contribution of such influences to phenotypic inequality by calculating the correlation between the phenotypes of genetically unrelated children reared together.

To the extent that environments also vary within families, this procedure will underestimate the overall impact of environmental inequality. But once again this danger will be obvious to all concerned. For policy purposes, moreover, one may not care about the overall impact of all conceivable forms of environmental inequality. It may be more useful to know the impact of specific sources of environmental variation. We may want to know, for example, how much phenotypic variation is attributable to variations in home environment. We cannot answer this question using twin data, since we do not know the correlation between a family's mean genotype and its mean on exogenous environmental variables. We could answer this question if we had data on unrelated children reared in the same homes.

(3) We can estimate $\underline{h}^2$ from twin samples only if we assume that the correlation between exogenous environmental influences on MZ twins is equal to the correlation for DZ twins. We cannnot test this assumption directly. We can, however, view the assumption as a special case of a more general hypothesis, namely that $r_{XX'}$ is the same for all sorts of twins reared in the same home. If this were the case, and if sex per se had no effect on a trait, we would expect mixed-sex twins' phenotypes to correlate as highly as same-sex DZ twins' phenotypes. If sex affected the phenotype under study, we would expect $r_{PP'\underline{D}} = r_{PP'\underline{D}''} + 2r^2_{PS}$ where r_{PS} is the correlation between the phenotype in question and sex. In fact, $r_{PP'\underline{D}} > r_{PP'\underline{D}} + 2r^2_{PS}$ for both test scores and education, though the difference is only significant for test scores. This means that $r_{XX'\underline{D}} > r_{XX'\underline{D}''}$, at least for test scores. This suggests that $r_{XX'\underline{M}}$ may also be greater than $r_{XX'\underline{D}}$. If so, our estimates of $\underline{h}^2$ are too high.

(4) Estimates of the correlation between genotype and phenotype also depend on the assumption that twins do not influence one another. We have only been able to test this assumption using one set of genes, namely those that determine sex. A male responent's educational attainment is higher if he has a male sibling, and it is higher yet if he has an identical male twin. Additional research is badly needed to see if this pattern holds for genes other than those that determine sex. Unfortunately, there is no immediate prospect that we will be able to identify an appreciable fraction of the relevant genes. One could make a start by looking at pairs of DZ twins or siblings in which one twin or sibling had a serious mental or physical disorder with a known genetic basis, while the other twin or sibling was normal. Meanwhile, the evidence on sex supports the notion of reciprocal influence, though not unambiguously. If reciprocal influence is important, our estmates of $\underline{h}^2$ are probably too high.

(5) When we began this research we had hoped to use the Project Talent twins and siblings to estimate the broad heritability of a host of traits, including not only test scores, educational attainment, and career plans but occupation and earnings. But the number of same-sex DZ pairs

was very small, and the number of pairs with followup data was even smaller. Worse yet, the same-sex DZ pairs with followup data were not altogether typical of the larger population from which they were drawn. These differences may be due to random sampling error, but the fact remains that they lead to serious problems when we try to compare same-sex DZ twins to MZ twins. Such comparisons yield excessively high point estimates of h^2 for career plans and implausibly low estimates of h^2 for test scores. The regression of educational differences on test score differences (b_Δ) for same-sex twins is also anomalous.

We had originally planned to reduce the standard error of our point estimates by pooling data on same-sex DZ twins with data on mixed-sex twins and siblings. This strategy would be legitimate if there were good reason to suppose that same-sex DZ twins, mixed-sex twins, and siblings were essentially similar. Unfortunately, our data strongly implied that $r_{XX'}$ was larger for same-sex twins than for mixed-sex twins or siblings. This ruled out pooling the samples.

(6) Our best point estimates of h^2 for both test scores and educational attainment are in the neighborhood of 0.50, though both estimates have large standard errors. Since test scores only explain about 30 percent of the variance in educational attainment, these results imply the genes have an appreciable direct effect on educational attainment, independent of test scores. This inference is supported by data from Taubman's sample.

(7) One can estimate the correlation between genotypes for any pair of outcomes using equation 84. Our point estimate of the correlation between genotypic test performance and genotypic educational attainment is 0.268. While this estimate may be biased downward and certainly has a very large standard error, it suggests that the genes affecting educational attainment are quite different from those affecting test performance. This finding suggests that we should not think of genetic advantages in one-dimensional terms. This conclusion is strongly supported by Taubman's twin data. Applying equation 84 to his 2000 pairs of twins, the implied correlations between genotypic educational attainment (G_B), genotypic occupational status (G_O), and genotypic earnings (G_Y) are $r_{GE,GX} = 0.57$, $r_{GE,GY} = 0.48$, $r_{GY,GU} = 0.53$.[22]

(8) Given the questionable assumptions required for extracting information from twin samples, conclusions based on such samples should be treated very cautiously unless or until they are confirmed using other samples as well. The most useful samples would cover genetically related individuals reared apart. But with the exception of half-siblings, such samples are hard to assemble without cooperation from adoption agencies, and such cooperation is difficult to obtain.[23] In the absence of data on relatives reared apart, our estimates could be substantially improved by collecting data on half-siblings and on adopted siblings reared together.

(9) From both a scholarly and a political viewpoint, the critical question is not how much variance genes can explain in a given outcome,

but how they explain it. Genes may affect an outcome by providing the basis for invidious racial or sexual discrimination, for example. Or they may affect an outcome by affecting the body's ability to metabolize sugar. Both the intellectual and practical implications of these two phenomena are totally different. If kinship research is to yield any useful results, we believe that it will require causal modeling. We have tried to illustrate this approach using data on IQ and education. Given our small and probably unrepresentative sample, our results should not be taken very seriously. They do, however, underline the importance of looking for mechanisms rather than just patterns of association.

FIGURE I

PATH MODEL OF RELATIONSHIPS BETWEEN PHENOTYPE (P), GENOTYPE (G), ENVIRONMENT (E), AND EXEGENOUS ENVIRONMENTAL FACTORS (X)

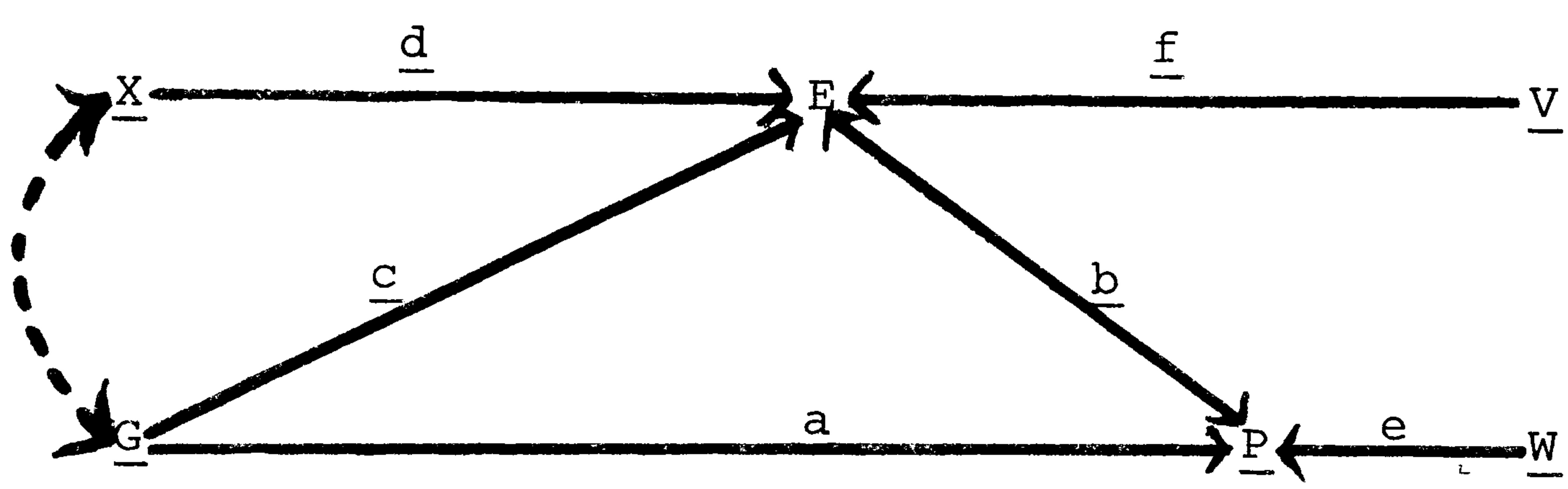


FIGURE 2

REDUCED-FORM MODEL OF DETERMINANTS OF PHENOTYPE

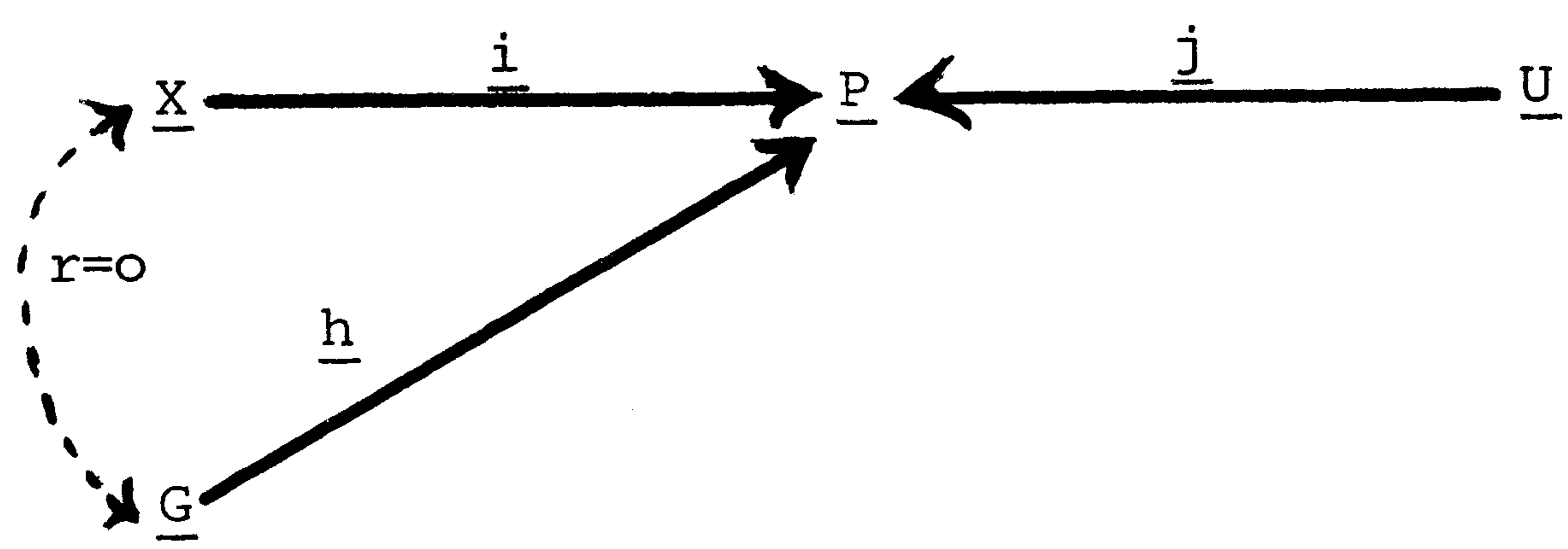


FIGURE 3

SOURCES OF PHENOTYPIC RESEMBLANCES BETWEEN TWINS REARED TOGETHER

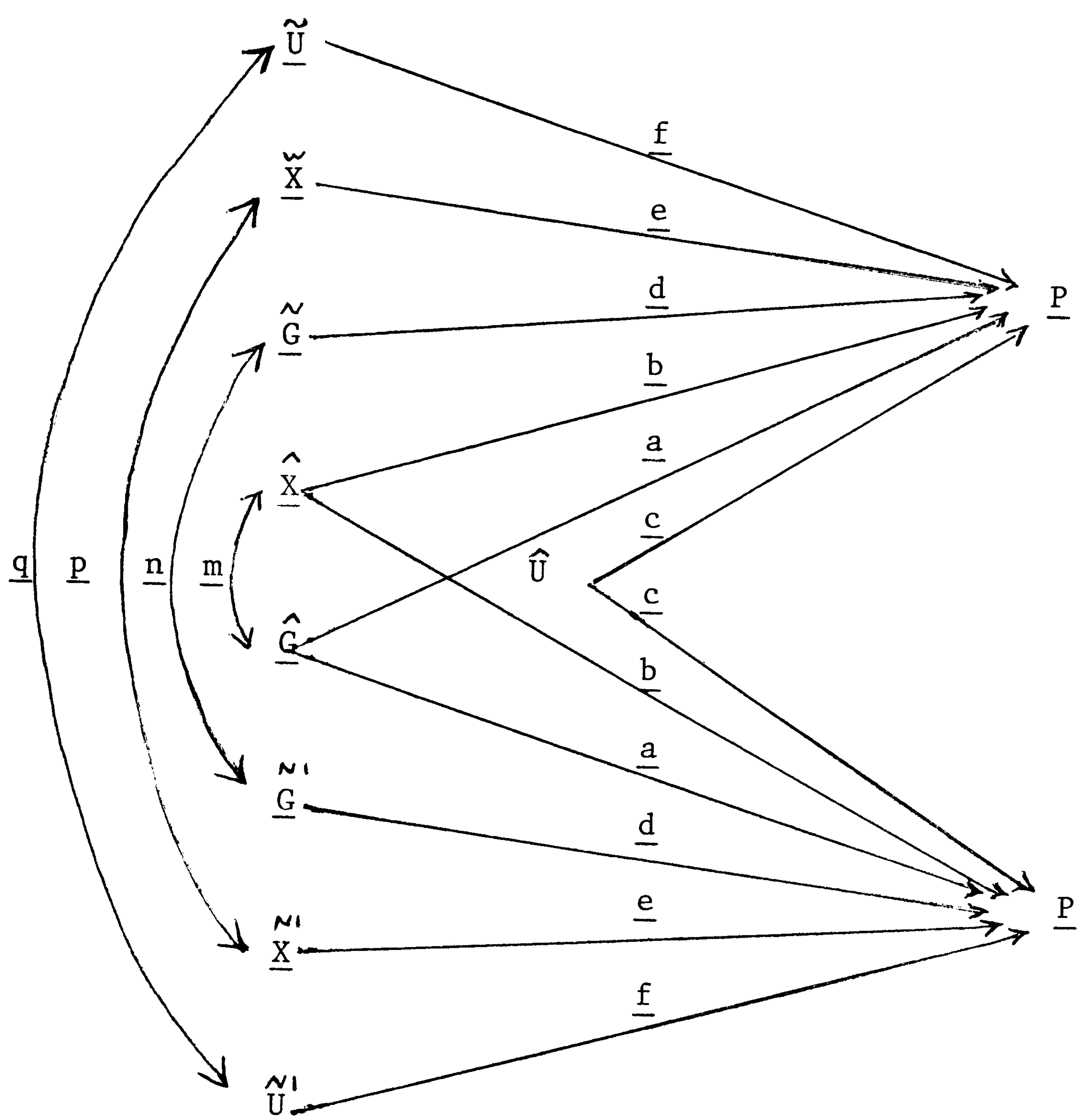
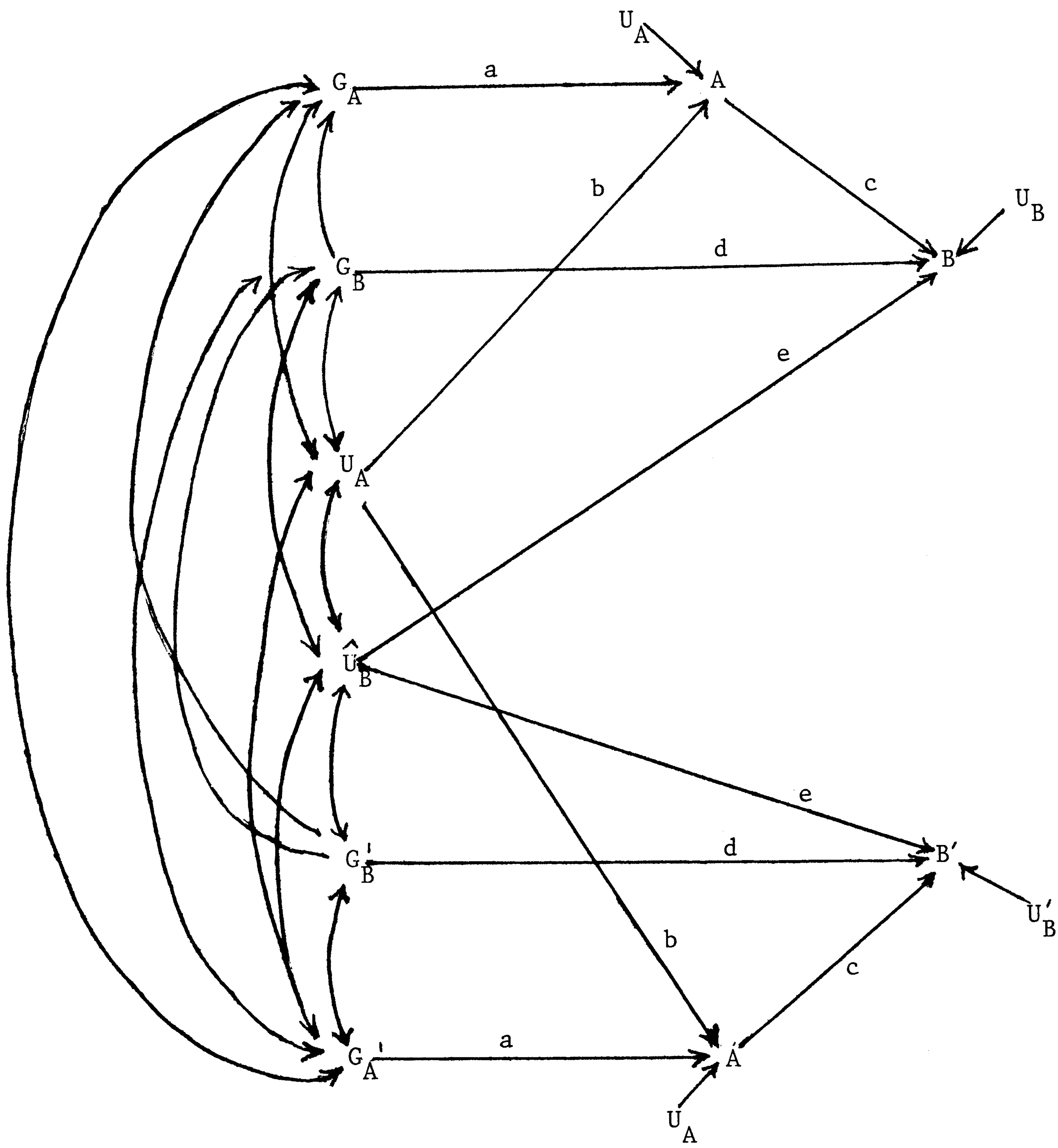


FIGURE 4

PATH MODEL OF THE SOURCES OF RESEMBLANCE BETWEEN TWINS' EDUCATIONAL ATTAINMENTS



FOOTNOTES

[1]We have used $\underline{h}^2$ as a symbol despite the fact that we are interested in "broad" heritability. Some investigators use $\underline{h}^2$ to symbolize "narrow" heritability, i.e. the ratio of the additive genetic variance to the total phenotypic variance. The narrow heritability is unknown in this example.

[2]This assumption is not quite as ridiculous as it sounds, since identical twins may compete with one another in ways that make their intrauterine environments dissimilar. This could offset the homogenizing influence of other factors on intrauterine environments.

[3]Unwary investigators often treat r_{MZA} as an estimate of $\underline{h}^2$. But $r_{MZA} = \underline{h}^2$ only if the total variance for separated twins is the same as the total variance for the population. If adopting parents provide children with more uniform environments than natural parents, or if random assignment reduces the extent to which genetically advantaged individuals are also environmentally advantaged, the variance for separated twins will be less than the population variance, and r_{MZA} will exceed $\underline{h}^2$. For a more detailed analysis see Jencks et al. (1972), Appendix A.

[4]See Project Talent (1972).

[5]Some of the spurious responses may have been due to random coding errors.

[6]Some time after retrieving the present sample, we used a computer routine to search Talent's files for 11th and 12th grade twins with 11-year followup data. We found 33 mixed-sex pairs that met Schoenfeldt's criteria for inclusion. Only 28 of these pairs appeared on his list. This implies that the full 1960 sample included (33/28) (1959) = 2309 pairs who met Schoenfeldt's criteria for inclusion. We did not scan for missing same-sex pairs, since we had no way of determining the zygosity of a pair unless Schoenfeldt had identified it and had sent the twins a zygosity questionnaire. Nor did we have Schoenfeldt's list of same-sex pairs of unknown zygosity, so such a scan would not have helped us assess his search procedure.

[7]Bulmer (1970).

[8]Shaycoft (1967) describes this 12 percent subsample in considerable

detail. For additional data see Jencks and Brown (1975).

[9]Because both students and keypunchers made errors in recording parents' first names, we did not require absolute agreement on spelling. We listed all near-matches. Most were clearly true matches. We excluded doubtful cases, on the grounds that inclusion of non-siblings would bias the results more than the exclusion of a few genuine pairs.

[10]Talent labels this composite C-002. Flanagan et al. (1964) describe it in more detail.

[11]These results are described in more detail in Jencks and Brown (1975). They are based on a Talent subsample retested in 12th grade. Retest scores are only available for a handful of twins and siblings.

[12]Because of the way Talent conducted its retesting program, one cannot compute composite C-002 scores at the time of retesting. For correlations between 9th and 12th grade scores on various subtests, see Shaycoft (1967) and Jencks and Brown (1975). Note that changes between 9th and 12th grade may be correlated across pairs of twins and siblings, making the intraclass correlation for twins tested at the same age higher than for siblings tested at different ages. The correlation between changes over time may be either genetic or non-genetic in origin. It could be genetic if, for example, changes between 9th and 12th grade depended on differing rates of physical maturation. It could be non-genetic if, for example, certain homes stressed mastering the higher school curriculum while others did not.

[13]For evidence that parental IQ exerts a positive effect on adopted children's IQ, see Jencks et al. (1972, Appendix A). Jencks et al. estimate r_{GX} at about 0.3, but see the correction in Loehlin et al. (1975). We have no analogous data for educational attainment.

[14]This is not strictly true, since certain genetic abnormalities such as Downs' syndrome are not inherited in the ordinary Mendelian sense but instead result from chromosomal damage. The risk of such abnormalities increases with maternal age and other environmental factors. Thus, if maternal age affected a given phenotype, r_{GX} could be non-zero. Such non-random events are not likely to explain much of the phenotypic variance, however, so we will ignore them.

[15]In principle, a 2x2x3 ANOVA has 12 cells and needs 11 dummies. But there are no male-female or female-male MZ pairs so we are left with 10 cells and 9 dummies.

[16]For exact formulae see Crow and Felsenstein (1967).

[17]For an explanation of why this is so, see Falconer (1960) or any other standard text on population genetics.

[18]For a bit of supporting evidence, see Eaves (1975).

[19]Jencks et al. (1972) estimated $\underline{h}^2$ at 0.45, but after correcting an error in Jencks et al.'s procedure for estimating covariance, Loehlin et al. (1975) obtained results implying that $\underline{h}^2 = 0.61$. When Eaves (1975) assumed no covariance between genotype and family environment, he got an implied value of 0.68 for $\underline{h}^2$.

[20]Taubman reports that his method of diagnosing zygosity was "almost 95 percent" accurate. So while we assume that $1 - E_{M*} - E_{D*} = 0.733 \pm 0.083$ in our Talent sample, we assume that $1 - E_{M*} - E_{D*} = 0.90$ for Taubman's sample. In the absence of evidence, we also assumed that $s^2_{EM*} = s^2_{ED*} = 0$. In fact, S^2_{EM*} and s^2_{ED*} must be positive, so Table 4 slightly understates the sampling error of $\underline{h}^2$ for Taubman's sample.

[21]For examples of this assumption see e.g. Burt (1943), Young and Gibson (1963), Eckland (1967), Herrnstein (1971), and Jencks et al. (1972).

[22]Computed from the correlation matrix in Behrman, Taubman, and Wales' chapter in this volume.

[23]J. M. Horn, et al (1976) are assembling IQ data on a new sample of adopted children in Texas -- the first since the early 1930's. Unfortunately, the variance of IQ in this sample is quite restricted, and no data are available on adult outcomes.

BIBLIOGRAPHY

Blau, Peter and Duncan, Otis Dudley. The American Occupational Structure. New York: Wiley, 1967.

Bulmer, M.G. The Biology of Twinning in Man. Oxford: Clarendon Press, 1970.

Burt, Cyril. "Ability and Income." British Journal of Educational Psychology, Vol. 13 (June 1943), pp. 83-98.

Carter-Saltzman, Louise and Scarr-Salapatek, Sandra. "Blood Group, Behavioral, and Morphological Differences Among Dizygotic Twins." University of Minnesota, 1975, xerox.

Crouse, James. "The Effects of Cognitive Skills on Economic Success", in Christopher Jencks, ed., Who Gets Ahead? Basic Books, forthcoming.

Crow, James and Felsenstein, J. "The Effect of Assortative Mating on the Genetic Composition of a Population." Eugenics Quarterly, Vol. 15 (June, 1968), pp. 85-97.

Duncan, Otis Dudley. "A Socioeconomic Index for all Occupations." in Albert J. Reiss, Occupations and Social Status, New York: Free Press, 1961.

Duncan, Otis Dudley; Featherman, David; and Duncan, Beverley. Socioeconomic Background and Achievement. New York: Seminar Press, 1972.

Eaglesfield, David. "The NORC Brothers Sample." in Christopher Jencks, ed., Who Gets Ahead? Basic Books, forthcoming.

Eaves, Lyndon J. "Testing Models for Variation in Intelligence." Heredity, Vol. 34, Part 1 (February 1975), pp. 132-136.

Eckland, Bruce. "Genetics and Sociology: A Reconsideration." American Sociological Review, Vol. 32 (April 1967), pp. 173-194.

Falconer, Douglas S. Introduction to Quantitative Genetics. New York: Ronald Press, 1960.

Flanagan, John C. et al. The American High School Student. Palo Alto: American Institutes for Research, 1964.

Goldberger, Arthur. "Statistical Inference in the Great IQ Debate." Madison: University of Wisconsin, Institute for Research on Poverty, 1975 (offset).

Goodman, Leo. "On the Exact Variance of Products," Journal of the American Statistical Association, Vol. 55 (December 1960), pp. 708-713.

Griliches, Zvi. "Estimating the Returns to Schooling: Some Econometric Problems." Cambridge: Harvard Institute of Economic Research, Discussion Paper 433, 1975.

Hauser, Robert and Featherman, David. "Equality of Schooling: Trends and Prospects." Sociology of Education, Vol. 49 (April 1976), pp. 99-120.

Herrnstein, Richard. "IQ," The Atlantic Monthly, September 1971, pp. 43-64.

Horn, J.M.; Loehlin, J.; and Willerman, L. "Texas Adoption Project." Behavior Genetics, Vol. 6 (1976), p. 109 (abstract).

Jensen, Arthur R. "Estimating the Limits of Heritability of Traits by Comparison of Monozygotic and Dizygotic Twins." Proceeedings of the National Academy of Science, Vol. 58 (1967), pp. 149-156.

Jencks, Christopher, et al. Inequality. New York: Basic Books, 1972.

Jencks, Christopher, and Brown, Marsha. "Effects of High Schools on Their Students." Harvard Educational Review, Vol. 45 (August 1975), pp. 273-324.

Loehlin, John; Lindzey, Gardner; and Spuhler, J.N. Race Differences in Intelligence. San Francisco: W.H. Freeman, 1975.

Loehlin, John and Nichols, Robert. Heredity, Environment, and Personality: A Study of 850 Sets of Twins, Austin: University of Texas Press, 1976.

Olneck, Michael. "The Kalamazoo Brothers' Sample." in Christopher Jencks, ed., Who Gets Ahead? Basic Books, forthcoming.

Project Talent. The Project Talent Data Bank: A Handbook. Palo Alto: American Institutes for Research, 1972.

Schoenfeldt, Lyle. "A Comparison of Two Analytic Procedures for Estimating Twin Zygosity." Palo Alto: American Institutes for Research, 1967 (offset).

__________. "Heritability-Environmental Components of the Project Talent Two-Day Test Battery." Measurement and Evaluation in Guidance Vol. 1 (Summer 1968), pp. 130-140.

Shaycoft, Marion. The High School Years: Growth in Cognitive Skills. Palo Alto: American Institutes for Research, 1967.

Young, Michael and Gibson, John. "In Search of an Explanation of Social Mobility." British Journal of Statistical Psychology, Vol. 16 (May 1963), pp. 27-36.

AN INSTRUMENTAL VARIABLE INTERPRETATION OF IDENTIFICATION IN VARIANCE-COMPONENTS AND MIMIC MODELS

Gary Chamberlain*
Harvard University

1. Introduction

Elsewhere I have attempted a general treatment of identification in variance-components models (Chamberlain [1976]). This paper is intended to clarify those results by giving them an instrumental variable (IV) interpretation. Although not all of the results fit easily into the IV framework, in many cases it gives the clearest indication of where the identification is coming from. Also it will often indicate quickly whether a new model is identified. Another advantage is that the IV identification is based on solving linear equations; hence there is no danger of multiple solutions.

The other new result is that our analysis of the variance-components model carries over directly to the Multiple-Indicator, Multiple-Cause Model (MIMIC) of Jöreskog and Goldberger. We extend their model by allowing the indicators to be jointly determined. Then we show that the identification conditions are identical to the conditions in the corresponding variance-components model. In the MIMIC model the observable "causes" substitute for the group structure of the variance-components model.

2. The Instrumental Variable Interpretation

One-Factor Model. We begin with the simplest version of the problem, a three-equation, one-factor model:

$$\begin{aligned} y_{1ij} &= && && + \varepsilon_{1ij} \\ y_{2ij} &= \gamma_{12} y_{1ij} && && + \varepsilon_{2ij} \\ y_{3ij} &= \gamma_{13} y_{1ij} && + \gamma_{23} y_{2ij} && + \varepsilon_{3ij}. \end{aligned}$$

The subscripts refer to the j^{th} individual in the i^{th} group. For example,

*I am grateful for discussions with A. Goldberger, Z. Griliches, J. Hausman, J. Heckman, J. Medoff, and P. Taubman. Financial Support was provided by the NIE and NSF.

the group could be a family, with observations on sibs within a family. We shall confine ourselves to triangular models in order to simplify the exposition and because these models are widely used with micro data on individuals. We shall also simplify the exposition by not carrying along any exogenous variables. This allows us to concentrate on the non-standard aspects of our identification results.

The availability of the group structure raises interesting possibilities for structuring the residuals. We assume that residuals from different groups are uncorrelated; but we still have to model the correlations within a group, both across equations and, for a given equation, across distinct members of the group.

A convenient way to model both types of correlation is to postulate latent variables connecting the residuals. A natural starting point is the one-factor model:

$$\varepsilon_{kij} = \lambda_k h_{ij} + \mu_{kij}, \qquad k = 1,2,3, \tag{2}$$

where the μ's are uncorrelated across equations (k) and across individual observations (j) within a group. So the correlations across equations are all generated by the common omitted h variable. The group structure of the residuals can be generated by assigning h a variance-components structure:

$$h_{ij} = f_i + g_{ij}, \tag{3}$$

where f_i is independent and identically distributed (i.i.d.) across groups, and g_{ij} is i.i.d. within groups and uncorrelated with f. Balestra and Nerlove (1966), Wallace and Hussein (1969), Nerlove (1971), Madalla (1971), and others have dealt with the single-equation version of this variance-components model.

A common criticism of these models is that the random effects are unlikely to be uncorrelated with the explanatory variables (see Mundlak [1976]). A good example, due to Z. Griliches, is a combined time-series, cross-section analysis of an agricultural production function. Then interpretations of the farm effects include soil quality and managerial efficiency. Assuming that the farms on better quality soil (or the better managed farms) are aware of this, then under competition with decreasing returns to the other inputs, those farms will use more of the other inputs. Hence the farm effects will be correlated with the explanatory variables.

More generally, the conventional random effects specification has two parts. The first part assumes that the marginal distribution of the effects is a random sample distribution. One way of judging this is deFinetti's (1937) exchangeability criterion. If the indices (ij) have no substantive content and are purely a labeling device, then permuting them will leave the form of the (subjective prior) distribution of the effects unchanged. In that case the distribution must be a random sample distribution. This criterion is most likely to be met in samples of individuals, for then there is often no natural ordering. Or in working within families, a two-level exchangeability may be appropriate; the random sample distribution of the f_i corresponds to exchangeability across families, and the random sample distribution of the g_{ij} corresponds to exchangeability within families.[1]

The second part of the random effects specification is a conditional distribution of the effects given the explanatory variables. Here the conventional assumption of independence is often inappropriate, as the production function example illustrated. But our model is not subject to this criticism, for we have built in explicitly the correlation between the random effects and the explanatory variables.

One application of our model is the estimation of earnings functions, controlling for an unobserved "ability" variable. The version of the model that is easiest to identify has y_1 = an early (pre-school) test score, y_2 = years of schooling, and y_3 = (log) earnings. Then if h is interpreted as early ability, it is plausible to exclude y_1 from the other equations ($\gamma_{12} = \gamma_{13} = 0$). For we interpret y_1 as measuring ability subject to an error (μ_1). The simplest interpretation of μ_1 is a pure test-retest error which is uncorrelated with everything but y_1. Then excluding y_1 from the other equations says that the measured score has no effect given the true score. This captures the standard errors-in-variables model.[2] In some cases it may be reasonable to include both the measured score and the true score—the measured score may have a credential or signalling effect. But then the model loses its errors-in-variables flavor and could be better classified as a general unobservables model.

With $\gamma_{12} = \gamma_{13} = 0$, the model can be identified by a simple instrumental variable (IV) argument, which shall be referred to as the proxy-IV approach. Use y_1 as a proxy for h in the y_3 equation:

$$y_3 = \gamma_{23}y_2 + \frac{\lambda_3}{\lambda_1} y_1 + \mu_3 - \frac{\lambda_3}{\lambda_1} \mu_1 . \tag{4}$$

Now we have an errors-in-variables problem. A solution is to use the value of y_1 for some other member of the group as an IV. Assume that there are at least two members per group, say $j = \alpha, \beta$. Then use $y_{1\alpha}$ as the proxy:

$$y_{3\alpha} = \gamma_{23}y_{2\alpha} + \frac{\lambda_3}{\lambda_1} y_{1\alpha} + \mu_{3\alpha} - \frac{\lambda_3}{\lambda_1} \mu_{1\alpha}, \tag{5}$$

and use $y_{1\beta}$ as an IV. This is legitimate since $y_{1\beta}$ is uncorrelated with $\mu_{3\alpha}$ and $\mu_{1\alpha}$, but it is correlated with $y_{1\alpha}$ as long as $\lambda_1 \sigma_f \neq 0$.

Next we shall examine a slight variation of this model, for which the simple proxy-IV argument does not hold. Suppose that the only restriction is $\gamma_{23} = 0$:

$$
\begin{aligned}
y_1 &= \qquad\qquad\quad \lambda_1 h + \mu_1 \\
y_2 &= \gamma_{12} y_1 + \lambda_2 h + \mu_2 \\
y_3 &= \gamma_{13} y_1 + \lambda_3 h + \mu_3 .
\end{aligned}
\tag{6}
$$

For example, let y_1 = years of schooling, y_2 = a late test, which depends on schooling, and y_3 = earnings. We interpret μ_2 as a test-retest error which is uncorrelated with everything except the test. The measured score is excluded from the third equation since the true score components--y_1 and h--are included.

Another application is for y_2 and y_3 to measure earnings in different years. Then h is the initial stock of human capital, which is augmented by schooling (y_1); μ_2 and μ_3 are transitory income; and the $\gamma_{23} = 0$ restriction says that there is no correlation between transitory income in the two years.

A related possibility is y_1 = schooling, y_2 = consumption, and y_3 = earnings, with h still interpreted as initial human capital. Then the $\gamma_{23} = 0$ restriction says that transitory income is uncorrelated with transitory consumption.

Under all three interpretations, the model is formally equivalent to the one used in Chamberlain and Griliches (1975). They examined directly the reduced form of the model, and showed that in general it is identified. Although straightforward, their approach is somewhat cumbersome. My initial efforts to find an IV interpretation were guided by the previous example. Hence I looked for a proxy variable for h, together with a suitable IV to treat the measurement error. But this approach does not work here. Instead the trick is to leave h in the residual and construct an IV that is uncorrelated with h. This will be referred to as the purged-IV approach.

It requires two stages to identify the γ's and a third stage to identify the λ's. The reduced form of the model is

$$
y_k = d_k h + e_k, \qquad k = 1,2,3, \tag{7}
$$

where

$$
\underset{\sim}{d} = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix} = \begin{bmatrix} \lambda_1 \\ \lambda_2 + \gamma_{12}\lambda_1 \\ \lambda_3 + \gamma_{13}\lambda_1 + \gamma_{23}(\lambda_2 + \gamma_{12}\lambda_1) \end{bmatrix},
$$

and

$$e = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 + \gamma_{12}\mu_1 \\ \mu_3 + \gamma_{13}\mu_1 + \gamma_{23}(\mu_2 + \gamma_{12}\mu_1) \end{bmatrix}.$$

The first stage is to use y_1 as a proxy for h in the reduced form equation for y_2:

$$y_2 = \frac{d_2}{d_1} y_1 + e_2 - \frac{d_2}{d_1} e_1. \tag{8}$$

The resulting errors-in-variables problem can be solved by using $y_{1\beta}$ as an IV for $y_{1\alpha}$, provided that $\lambda_1 \sigma_f \neq 0$. Then form the residual, thereby purging y_2 of its dependence on h:

$$z = y_2 - \frac{d_2}{d_1} y_1 = e_2 - \frac{d_2}{d_1} e_1. \tag{9}$$

The second stage is to use z as an IV for y_1 in the structural y_3 equation in (6). z is an appropriate IV since it is uncorrelated with h and with μ_3, but it is correlated with y_1 if $\lambda_2 E(\mu_1^2) \neq 0$. So we require that h appear directly in the y_2 equation, and that y_1 is not proportional to h—otherwise we could never separate the effects of y_1 and h. In order to identify γ_{12}, we can interchange the y_2 and y_3 equations and repeat the two stages.

With γ_{13} identified, in the third stage we form the residual

$$w = y_3 - \gamma_{13} y_1 = \lambda_3 h + \mu_3. \tag{10}$$

Then use y_1 as a proxy for h:

$$w = \frac{\lambda_3}{\lambda_1} y_1 + \mu_3 - \frac{\lambda_3}{\lambda_1} \mu_1. \tag{11}$$

Now $\frac{\lambda_3}{\lambda_1}$ can be identified by a third application of instrumental variables, using $y_{1\beta}$ as an IV for $y_{1\alpha}$. (Note that only the ratio of the λ's is identified, due to the indeterminate scale of the latent variable.) Interchanging the second and third equations, we can identify λ_2/λ_1 by repeating this procedure.

Extension to Larger Systems. In larger one-factor models it is still true that identification requires at least one restriction in addition to those implied by the triangular structure. Say that the restriction excludes y_2 from the t^{th} equation. The reduced form is $y_k = d_k h + e_k$. Using y_1 as a proxy gives

$$y_k = \frac{d_k}{d_1} y_1 + e_k - \frac{d_k}{d_1} e_1, \quad k = 2, \ldots, t-1. \tag{12}$$

Then d_k/d_1 is identified by using $y_{1\beta}$ as an IV for $y_{1\alpha}$. Given d_k/d_1, we can form the residuals

$$z_k = y_k - \frac{d_k}{d_1} y_1 = e_k - \frac{d_k}{d_1} e_1, \quad k = 2, \ldots, t-1. \tag{13}$$

Now there are t-2 IV's for the t-2 variables that remain on the right hand side of the t^{th} equation after y_s has been excluded.[3]

Having identified the coefficients in the t^{th} equation, we form the residual

$$y_t^* = y_t - \gamma_{1t} y_1 - \cdots - \gamma_{t-1,t}\, y_{t-1} = \lambda_t h + \mu_t. \tag{14}$$

Then we can identify the k^{th} equation, with $k < t$, by using $y_1, \ldots, y_{k-1}$ together with y_t^* to form k-1 purged IV's. Note that we must use y_t^* and not y_t since (in general) y_t is correlated with μ_k.

The identification of the equations that follow y_t is not, however, as straightforward. We can illustrate by adding an equation to our three equation example:

$$y_4 = \gamma_{14} y_1 + \gamma_{24} y_2 + \gamma_{34} y_3 + \lambda_4 h + \mu_4. \tag{15}$$

Assuming that the previous equations have been identified, we can replace y_2 by $y_2^* = y_2 - \gamma_{12} y_1$, and replace y_3 by $y_3^* = y_3 - \gamma_{13} y_1 - \gamma_{23} y_2$:

$$y_4 = \gamma_{14}^* y_1 + \gamma_{24}^* y_2^* + \gamma_{34} y_3^* + \lambda_4 h + \mu_4, \tag{16}$$

where $\gamma_{24}^* = \gamma_{24} + \gamma_{34}\gamma_{23}$ and $\gamma_{14}^* = \gamma_{14} + \gamma_{34}\gamma_{13} + \gamma_{24}^*\gamma_{12}$.

Note that γ_{34} is unaffected by this transformation. Now use y_1 as a proxy for h:

$$y_4 = \gamma_{24}^* y_2^* + \gamma_{34} y_3^* + \pi_1 y_1 + \mu_4 - \frac{\lambda_4}{\lambda_1} \mu_1, \tag{17}$$

where $\pi_1 = \gamma^*_{14} + \dfrac{\lambda_4}{\lambda_1}$. Since μ_1 is uncorrelated with y^*_2 and y^*_3, we can use $y^*_{2\alpha}$ and $y^*_{3\alpha}$ together with $y_{1\beta}$ as instrumental variables. This will identify γ_{34}, allowing us to form $\tilde{y}_4 = y_4 - \gamma_{34} y_3$. Then the $\tilde{y}_4$ equation excludes y_3, and so we can use y_1, y_2, and y_3 to form two purged-IV's for y_1 and y_2, thereby identifying γ_{14} and γ_{24}.

The generalization to additional equations proceeds in a similar fashion. We rewrite the fifth equation in terms of y_1, y^*_2, y^*_3, and y^*_4. The coefficient of $y_4(\gamma_{45})$ is unaffected by this. Then with y_1 as a proxy for h, only y_1 requires a cross-member IV since μ_1 is uncorrelated with the y^*_j, $j = 2,3,4$. Thus we can identify γ_{45} and form $\tilde{y}_5 = y_5 - \gamma_{45} y_4$. Now the $\tilde{y}_5$ equation has an additional restriction, and so it can be identified by the purged-IV technique.

Some Pitfalls. Before treating the multi-factor case, I want to point out some possible misuses of instrumental variables. Going back to our first example, say that $\gamma_{13} = 0$ but $\gamma_{12} \neq 0$. As before, use y_1 as a proxy for h in the y_3 equation:

$$y_3 = \gamma_{23} y_2 + \frac{\lambda_3}{\lambda_1} y_1 + \mu_3 - \frac{\lambda_3}{\lambda_1}\mu_1. \qquad (18)$$

Since y_2 is now correlated with μ_1, instrumental variables are needed for both $y_{1\alpha}$ and $y_{2\alpha}$. A possibility is to use $y_{1\beta}$ and $y_{2\beta}$. But then the IV normal equations require the inversion of a singular matrix (note that we are suppressing the intercepts and assuming that all means are zero):

$$E\begin{bmatrix} y_{1\beta}\, y_{1\alpha} & y_{1\beta}\, y_{2\alpha} \\ y_{2\beta}\, y_{1\alpha} & y_{2\beta}\, y_{2\alpha} \end{bmatrix} = \begin{bmatrix} d_1^2 & d_1 d_2 \\ d_1 d_2 & d_2^2 \end{bmatrix} \sigma_f^2, \qquad (19)$$

where the d's are the reduced form coefficients of h. So this misuse of instrumental variables violates a rank condition.[4]

A related pitfall could arise if there were three or more members per group. Then we might consider using, say, $y_{1\beta}$ and $y_{1\delta}$ as IV's for $y_{1\alpha}$ and $y_{2\alpha}$ in (18). But once again the IV normal equations would be singular:

$$E\begin{bmatrix} y_{1\beta}\, y_{1\alpha} & y_{1\beta}\, y_{2\alpha} \\ & \\ y_{1\delta}\, y_{1\alpha} & y_{1\delta}\, y_{2\alpha} \end{bmatrix} = \begin{bmatrix} d_1^2 & d_1 d_2 \\ & \\ d_1^2 & d_1 d_2 \end{bmatrix} \sigma^2_f. \qquad (20)$$

Our second, more novel, application of instrumental variables also has a potential pitfall. Say that there are no restrictions on the γ 's in the three-equation model. Then in order to follow the strategy of allowing h to remain in the residual, in the third equation we need IV's for y_1 and y_2 that are uncorrelated with h. We have already indicated how to purge y_2 of its dependence on h to form $z = y_2 - \frac{d_2}{d_1} y_1$. A similar procedure can be applied to y_1. We use y_2 as a proxy for h, with $y_{2\beta}$ as an IV for $y_{2\alpha}$. Then we form the residual $z^* = y_1 - \frac{d_1}{d_2} y_2$. This z* meets the criterion of being uncorrelated with h and μ_3. But $z^* = -\frac{d_1}{d_2} z$, and so an attempt to use both z and z* as IV's will fail the rank condition. Although obvious here, this error can take a more subtle form in the multi-factor models.

Multi-factor Models: Errors-in-Variables. We shall begin with an extension of the one-factor errors-in-variables model. The key property of an errors-in-variables model is that at least one variable is excluded from all of the other equations. Consider the following two-factor model:

$$\begin{aligned} y_1 &= \lambda_{11} h_1 + \lambda_{21} h_2 + \mu_1 \\ y_2 &= \lambda_{12} h_1 + \lambda_{22} h_2 + \mu_2 \\ y_3 &= \gamma_{23} y_2 + \lambda_{13} h_1 + \lambda_{23} h_2 + \mu_3 \\ y_4 &= \gamma_{24} y_2 + \gamma_{34} y_3 + \lambda_{14} h_1 + \lambda_{24} h_2 + \mu_4. \end{aligned} \qquad (21)$$

Since y_1 is excluded from the other equations, we can use it as a proxy variable:

$$y_1 = h_1^* + \mu_1, \qquad (22)$$

where $h_1^* = \lambda_{11} h_1 + \lambda_{21} h_2$. Then rewrite the factor structure in the other equations as

$$\frac{\lambda_{1j}}{\lambda_{11}} h_1^* + \lambda_{2j}^* h_2, \text{ with } \lambda_{2j}^* = \lambda_{2j} - \lambda_{1j} \frac{\lambda_{21}}{\lambda_{11}}, \; j = 2,3,4, \text{ and}$$

$\lambda_{21}^* = 0$. This illustrates how easy it is to transform the factor structure and indicates that the λ's will not be identified unless there are prior restrictions on them.

If $\gamma_{23} = \gamma_{24} = 0$ also, then we can use y_2 as a proxy variable:

$$y_2 = h_2^* + \mu_2, \tag{23}$$

where $h_2^* = \lambda_{12} h_1 + \lambda_{22} h_2$. Then rewrite the factor structure of the other equations using h_1^* and h_2^*.[5] The objective is simply to give the problem a more standard errors-in-variables appearance. The important point is that some combination of y_1 and y_2 will measure $\lambda_{14} h_1 + \lambda_{24} h_2$ subject to an error that is a combination of μ_1 and μ_2.

Using y_1 and y_2 as proxies in the y_4 equation gives (with $\gamma_{23} = \gamma_{24} = 0$):

$$y_4 = \gamma_{34} y_3 + \lambda_{14}^* y_1 + \lambda_{24}^* y_2 + \mu_4 - \lambda_{14}^* \mu_1 - \lambda_{24}^* \mu_2. \tag{24}$$

The errors-in-variables problem with y_1 and y_2 can be treated by using $y_{1\beta}$ and $y_{2\beta}$ as IV's for $y_{1\alpha}$ and $y_{2\alpha}$. Since the y_3 equation excludes y_1 and y_2, we can use $y_{3\alpha}$ as an IV for itself. If we had to use $y_{3\beta}$, then the rank condition would fail, for the rank of the cross-member covariance matrix ($E[\underset{\sim}{y}_\alpha \underset{\sim}{y}_\beta']$) cannot exceed the number of group factors.[6]

Now suppose that y_2 is not excluded from the y_3 and y_4 equations. For example, we could have y_1 = an early test score; y_2 = years of schooling; y_3 = a late test score, which depends on schooling; and y_4 = earnings, which also depends on schooling. The factors h_1 and h_2 are interpreted as different types of "ability," with μ_1 and μ_3 interpreted as measurement errors. Then a plausible additional restriction is $\gamma_{34} = 0$. This excludes the late test score from the earnings equation ($\gamma_{14} = 0$ excludes the early score) since the true score components—y_2, h_1, and h_2—are included. Estimates of such a two-factor model are presented in

Chamberlain and Griliches (1976).

This model is a combination of the two previous one-factor examples. y_1 can be used as a proxy for h_1^*, but there is no direct proxy for h_2^*, suggesting that one additional restriction is needed. This is correct, and we shall impose $\gamma_{34} = 0$. The appropriate strategy, however, is not to use y_1 as a proxy in the y_4 equation. Rather we follow the purged-IV approach, allowing h_1 and h_2 to remain in the y_4 residual and purging y_2 of its dependence on the factors.

First set up the reduced form for y_1, y_2, and y_3:

$$y_1 = h_1^* + \mu_1 \tag{25}$$

$$y_2 = h_2^* + \mu_2$$

$$y_3 = \lambda_{13}^* h_1^* + (\lambda_{23}^* + \gamma_{23}) h_2^* + \mu_3 + \gamma_{23}\mu_2.$$

Then use y_1 and y_2 as proxies for h_1^* and h_2^* in the y_3 equation:

$$y_3 = \lambda_{13}^* y_1 + (\lambda_{23}^* + \gamma_{23}) y_2 + \mu_3 - \lambda_{13}^* \mu_1 - \lambda_{23}^* \mu_2 . \tag{26}$$

This equation can be identified by using $y_{1\beta}$ and $y_{2\beta}$ as IV's for $y_{1\alpha}$ and $y_{2\alpha}$. It is permissible to use two cross-member IV's since this is a two-factor model.[7] Then form the purged residual:

$$z = y_3 - \lambda_{13}^* y_1 - (\lambda_{23}^* + \gamma_{23}) y_2 =$$

$$\mu_3 - \lambda_{13}^* \mu_1 - \lambda_{23}^* \mu_2. \tag{27}$$

Now z can be used as an IV for y_2 in the y_4 equation. It is uncorrelated with h_1, h_2, and μ_4; but it is correlated with y_2 if $\lambda_{23}^* E(\mu_2^2) \neq 0$. $\lambda_{23}^* \neq 0$ requires that h_2^* enter directly the y_3 equation; equivalently, it requires that $\lambda_{11}/\lambda_{21} \neq \lambda_{13}/\lambda_{23}$, so that the factor coefficients in the first and third equations are not proportional to each other. In addition, $E(\mu_2^2) \neq 0$ requires variation in y_2 that is uncorrelated with h_1 and h_2.

Interchanging the y_3 and y_4 equations, we can identify γ_{23} by repeating this purged-IV approach. The identification of the rest of the model, however, it not always so easy. As in the one-factor model, the problem is with the equations that follow the one with the extra restriction. We can illustrate by adding a fifth equation to the model (the proxy variable y_1 is still excluded from all of the equations):

$$y_5 = \gamma_{25} y_2 + \gamma_{35} y_3 + \gamma_{45} y_4 + \lambda^*_{15} h^*_1 + \lambda^*_{25} h^*_2 + \mu_5. \tag{28}$$

First we identify the γ's in the third and fourth equations (as above) and form the residuals:

$$y^*_k = \lambda^*_{1k} h^*_1 + \lambda^*_{2k} h^*_2 + \mu_k, \qquad k = 3,4. \tag{29}$$

Then rewrite the y_5 equation using the y^*'s:

$$y_5 = \gamma^*_{25} y_2 + \gamma^*_{35} y^*_3 + \gamma_{45} y^*_4 + \lambda^*_{15} h^*_1 + \lambda^*_{25} h^*_2 + \mu_5. \tag{30}$$

Note that the γ_{45} coefficient is not affected by this transformation.[8] Use y_1 and y_2 as proxies for h^*_1 and h^*_2:

$$y_5 = \gamma^*_{35} y^*_3 + \gamma_{45} y^*_4 + \lambda^*_{15} y_1 + (\gamma^*_{25} + \lambda^*_{25}) y_2 + \mu_5 - \lambda^*_{15} \mu_1 - \lambda^*_{25} \mu_2. \tag{31}$$

Since μ_1, μ_2, and μ_5 are uncorrelated with y^*_3 and y^*_4, we can use $y^*_{3\alpha}$ and $y^*_{4\alpha}$ as IV's together with $y_{1\beta}$ and $y_{2\beta}$. Thus γ_{45} is identified, and so we can form $\tilde{y}_5 = y_5 - \gamma_{45} y_4$. Then the $\tilde{y}_5$ equation excludes y_4 (and y_1); hence we can form two purged-IV's from $y_1, \ldots, y_4$ in order to identify the two remaining parameters in the y_5 equation, γ_{25} and γ_{35}.

Another Multi-Factor Example. The errors-in-variables models required one restriction in addition to the proxy exclusions. In the four-equation, two-factor model there were four restrictions. In fact, depending on their placement, only three restrictions may be needed.[9] For example, consider the following model:

$$y_1 = \lambda_{11} h_1 + \lambda_{21} h_2 + \mu_1 \tag{32}$$

$$y_2 = \lambda_{12} h_1 + \lambda_{22} h_2 + \mu_2$$

$$y_3 = \gamma_{23} y_2 + \lambda_{13} h_1 + \lambda_{23} h_2 + \mu_3.$$

$$y_4 = \gamma_{14} y_1 + \gamma_{34} y_3 + \lambda_{14} h_1 + \lambda_{24} h_2 + \mu_4.$$

This model is interesting because it requires a combination of the proxy-IV and the purged-IV approaches. In this respect the identification problem is similar to the one presented by the equations following the restricted equation in the previous examples.

The trick is to use y_2 and y_3 as proxies for h_1 and h_2:

$$y_4 = \gamma_{14} y_1 + \pi_1 y_2 + \pi_2 y_3 + \mu_4 + \psi_1 \mu_2 + \psi_2 \mu_3. \tag{33}$$

(π_2 includes γ_{34}.) Since y_1 is uncorrelated with μ_2, μ_3, and μ_4, we can use $y_{1\alpha}$ as an IV, along with $y_{2\beta}$ and $y_{3\beta}$. (The exclusion of y_1 from the y_2 and y_3 equations is necessary, for otherwise the errors in y_2 and y_3 as proxies for h_1 and h_2 would be correlated with y_1, requiring a cross-member IV for y_1. But we are allowed only two such IV's in the two-factor model.) So we can identify γ_{14} and form the residual $\tilde{y}_4 = y_4 - \gamma_{14} y_1$. Now the y_4 equation excludes y_1 (and y_2); hence, we can use y_1, y_2, and y_3 to form a purged IV for the remaining variable, y_3.

3. The MIMIC Model

The following is an example of the Jöreskog-Goldberger Multiple-Indicator, Multiple-Cause (MIMIC) model:

$$\begin{aligned} y_{1i} &= \lambda_1 h_i + \mu_{1i} \\ y_{2i} &= \lambda_2 h_i + \mu_{2i} \\ y_{3i} &= \lambda_3 h_i + \mu_{3i}, \end{aligned} \tag{34}$$

where

$$h_i = \underset{\sim}{x}'_i \underset{\sim}{\beta} + f_i.$$

The unobservable h is partitioned into a projection onto observable x's (the causes) and a residual (f) that is uncorrelated with $\underset{\sim}{x}$ by construction. The y's are indicators of h, with residuals μ that are uncorrelated with f and with each other.

We shall extend the MIMIC model by allowing the indicators to depend upon each other in a recursive fashion:

$$\begin{aligned} y_{1i} &= && + \lambda_1 h_i + \mu_{1i} \\ y_{2i} &= \gamma_{12} y_{1i} && + \lambda_2 h_i + \mu_{2i} \\ y_{3i} &= \gamma_{13} y_{1i} + \gamma_{23} y_{2i} && + \lambda_3 h_i + \mu_{3i}\,. \end{aligned} \tag{35}$$

This is our one-factor model except that there is no grouping device, and so h has only a single subscript. We shall see, however, that the observable x's can substitute for the group structure. For example, in the models of income determination, $\underset{\sim}{x}$ could consist of observable family background variables.

Consider first our original errors-in-variables example, with $\gamma_{12} = \gamma_{13} = 0$. Then as before we use y_1 as a proxy for h:

$$y_3 = \gamma_{23} y_2 + \frac{\lambda_3}{\lambda_1} y_1 + \mu_3 - \frac{\lambda_3}{\lambda_1} \mu_1. \tag{36}$$

The measurement error in y_1 is treated by using $\underset{\sim}{x}$ (or rather some linear combination of the x's) as an IV for y_1. y_2 can be used as an IV for itself. So the solution is identical to the variance-components case, except that the instrumental variable is based on $\underset{\sim}{x}$ instead of using y_1 for some other member of the group.

Note that if $\gamma_{12} \neq 0$ then y_2 is correlated with μ_1 and cannot be used as an IV. Then it is tempting to use the x's as IV's for y_1 and y_2, provided that there are at least two variables in $\underset{\sim}{x}$. But y_1 and y_2 depend on the same linear combination of the x's (up to a scale factor); hence the IV normal equations do not have full rank.

Consider next our second one-factor example, in which only $\gamma_{23} = 0$. As before we construct a purged IV by using y_1 as a proxy for h in the y_2 equation:

$$y_2 = \frac{d_2}{d_1} y_1 + e_2 - \frac{d_2}{d_1} e_1 , \tag{37}$$

where $d_1 = \lambda_1$, $d_2 = \lambda_2 + \gamma_{12} \lambda_1$, $e_1 = \mu_1$, and $e_2 = \mu_2 + \gamma_{12} \mu_1$. We identify d_2/d_1 by using $\underset{\sim}{x}$ as an IV for y_1. Then form the residual

$$z = y_2 - \frac{d_2}{d_1} y_1 = e_2 - \frac{d_2}{d_1} e_1. \tag{38}$$

z is a suitable IV for y_1 in the y_3 equation. So the approach is identical to the one used in the variance-components model, except that $\underset{\sim}{x}$ is used to form an IV for the proxy. Once the proxy coefficient $(d_2/\tilde{d}_1)$ has been identified, the remaining steps are exactly the same.[10]

For a multi-factor example, consider the one-proxy, two-factor model:

$$\begin{aligned} y_1 &= \lambda_{11} h_1 + \lambda_{21} h_2 + \mu_1 \\ y_2 &= \lambda_{12} h_1 + \lambda_{22} h_2 + \mu_2 \\ y_3 &= \gamma_{23} y_2 + \lambda_{13} h_1 + \lambda_{23} h_2 + \mu_3 \\ y_4 &= \gamma_{24} y_2 + \lambda_{14} h_1 + \lambda_{24} h_2 + \mu_4 \end{aligned} \tag{39}$$

Instead of a variance-components structure for the factors, we have

$$\begin{aligned} h_{1i} &= \underset{\sim}{x}'_i \underset{\sim}{\beta}_1 + f_{1i} \\ h_{2i} &= \underset{\sim}{x}'_i \underset{\sim}{\beta}_2 + f_{2i} . \end{aligned} \tag{40}$$

As before we rewrite the factor structure so that $y_1 = h_1^* + \mu_1$ and $y_2 = h_2^* + \mu_2$; then use y_1 and y_2 as proxies for h_1^* and h_2^* in the y_3 equation:

$$y_3 = \lambda_{13}^* y_1 + (\lambda_{23}^* + \gamma_{23}) y_2 + \mu_3 - \lambda_{13}^* \mu_1 - \lambda_{23}^* \mu_2 . \tag{41}$$

This equation can be identified by using the variables in $\underset{\sim}{x}$ as instrumental variables. If the IV normal equations are to have full rank, there must be at least two variables in $\underset{\sim}{x}$, and $\underset{\sim}{\beta}_1$ must not be proportional to $\underset{\sim}{\beta}_2$. Then we can form the residual $z = y_3 - \lambda_{13}^* y_1 - (\lambda_{23}^* + \gamma_{23}) y_2$ and use z as an IV for y_2 in the y_4 equation. Interchanging the y_2 and y_4 equations we can identify γ_{23} by repeating this purged-IV approach. The identification of additional equations follows the procedure that was used in the variance-components case.[11]

4. Summary and Extensions

We have provided an instrumental variable interpretation for a number of results on the identification of variance-components models. The IV interpretation clarifies where the identification is coming from. Also an appropriate combination of our proxy-IV and purged-IV techniques will often indicate quickly if a new model is identified. Another advantage is that the IV identification is based on solving linear equations; hence there is no danger of multiple, locally isolated solutions.

The other principal new result is that our analysis of the variance-components model carries over directly to a simultaneous equations version of the MIMIC model. The general case is a simultaneous equations model in which the residuals have a factor analytic structure. Each factor is partly determined by a set of exogenous variables. We have concentrated on the triangular case, but the correspondence between the two models holds in general.

As for future work, it will be useful to extend the results to models that include both discrete and continuous endogenous variables. Consider a triangular model relating y_1, y_2, and y_3, where y_1 and y_3 are continuous, but y_2 takes on only the values zero or one. In addition, there is an omitted (continuous) random variable h that affects all the y's. Then following Lazarsfeld (1968) and Heckman and Willis (1975, 1977), we might use the following model:

$$
\begin{aligned}
y_1 &= \lambda_1 h + \mu_1 \\
\tilde{y}_2 &= \gamma_{12} y_1 + \lambda_2 h + \mu_2 \\
y_3 &= \gamma_{13} y_1 + \gamma_{23} y_2 + \lambda_3 h + \mu_3 ,
\end{aligned}
\tag{42}
$$

where $\tilde{y}_2$ is an unobservable continuous random variable such that $y_2 = 1$ if $\tilde{y}_2 \geq 0$ and $y_2 = 0$ if $\tilde{y}_2 < 0$. It seems plausible that the identification conditions for the continuous case, when $\tilde{y}_2 = y_2$ is observable, will suffice to identify the model. In fact, this type of model may be identifiable under weaker conditions, by exploiting non-normality.

For example, consider the following MIMIC model:

$$
\begin{aligned}
h_i &= x_i \beta + f_i \\
\tilde{y}_{1i} &= \lambda_1 h_i + \mu_{1i} \\
y_{2i} &= \gamma_{12} y_{1i} + \lambda_2 h_i + \mu_{2i},
\end{aligned}
\tag{43}
$$

where $y_1 = 1$ if $\tilde{y}_1 \geq 0$ and $\tilde{y}_1 = 0$ if $y_1 < 0$. As an application, say that h is an unobserved ability variable, x is a family background variable such as family wealth, y_1 is a dummy variable for college completion, and y_2 is earnings. This one-factor model is not identified from the covariances, since one restriction is needed in addition to the triangular structure. We shall see, however, that it is identified from higher order moments.

The regression function is

$$
E(y_2 | x) = \gamma_{12} E(y_1 | x) + x\beta\lambda_2 . \tag{44}
$$

If f and μ_1 have normal distributions, then

$$
E(y_1 | x) = \mathrm{Prob}(\tilde{y}_1 \geq 0 | x) = \Phi(x \eta), \tag{45}
$$

where Φ is the standard normal cumulative function and

$\eta = \lambda_1 \beta/(\lambda_1^2 \sigma_f^2 + \sigma_1^2)^{1/2}$. η can be identified by a probit analysis of y_1 conditional on x. Then we have

$$E(y_2 | x) = x\beta\lambda_2 + \gamma_{12} \Phi(x \eta); \tag{46}$$

thus $\beta\lambda_2$ and γ_{12} are identified by the linear regression of y_2 on x and $\Phi(x \eta)$.[12]

If the original specification had allowed background to have a direct effect on college completion and on earnings, so that

$$\begin{aligned} \tilde{y}_1 &= x \delta \quad + \lambda_1 h + \mu_1 \\ y_2 &= x \pi + \gamma_{12} y_1 + \lambda_2 h + \mu_2, \end{aligned} \tag{47}$$

then η would be $(\delta + \lambda_1 \beta)/(\lambda_1^2 \sigma_f^2 + \sigma_1^2)^{1/2}$ and the coefficient of x in the regression of y_2 on x and $\Phi(x \eta)$ would be $\pi + \beta\lambda_2$ instead of $\beta \lambda_2$. But the coefficient of $\Phi(x \eta)$ in that regression would still identify γ_{12}.

There is a simple instrumental variable interpretation of these results, which helps to assess their usefulness. Let

$$y_2 = x(\pi + \beta \lambda_2) + \gamma_{12} y_1 + \lambda_2 f + \mu_2. \tag{48}$$

Now use x and x^2 as instrumental variables for x and y_1. Stated in this way the procedure seems rather contrived, for we can always manufacture additional instrumental variables from higher order terms.

There are two objections to such a procedure. One is that the partial correlation between y_1 and x^2, partialling on x, may be zero. This objection is less valid when y_1 is binary, for then its regression on x is unlikely to be linear. The second objection relates to the exclusion of x^2 from the y_2 equation. This exclusion requires that the direct effect of x on earnings is linear, and that the regression of h on x is linear. We are often satisfied with linear specifications as reasonable approximations; it seems questionable, however, to let the identification of the model rest solely on the exclusion of x^2 and higher order terms. Nevertheless, it will be useful to make comparisons with our other approach, which does not rely on higher order moments for identification.

FOOTNOTES

[1]We are abstracting from birth-order effects.

[2]This sort of model has been dealt with by Wiley (1973). For errors-in-variables models without a group structure, see Geraci and Goldberger (1971) and Geraci (1974, 1975).

[3]Chamberlain (1976) derives the following condition to ensure that the IV normal equations have full rank:

$$\sigma_f \lambda_t(\lambda_s/\sigma_s^2 - \sum_{k=s+1}^{t-1} \gamma_{sk} \lambda_k/\sigma_k^2) \neq 0,$$

where $\sigma_k^2 = E(\mu_k^2)$. In order to interpret this rank condition, we shall say that equation k is connected to the rest of the structure if $\lambda_k \neq 0$. Then the condition requires that the exclusion occur in a connected equation. If $\lambda_s = 0$, so that the excluded variable (y_s) is actually exogenous, then y_s must appear (with a non-zero coefficient) in a connected equation preceding the one it is excluded from. This condition is almost surely satisfied if there are no a priori constraints on the factor coefficients.

[4]The problem is that the covariances across different members of a group are all generated by a single common factor (f); hence no cross-member covariance matrix can have a rank greater than one.

[5]This transformation of the factor structure requires that $\lambda_{11}/\lambda_{21} \neq \lambda_{12}/\lambda_{22}$, so that the factor coefficients in the first and second equations are not proportional to each other.

[6]Note that the identification argument is unaffected if $\gamma_{12} \neq 0$. This is equivalent to allowing the measurement errors in the two proxies to be correlated with each other.

[7]We assume, of course, that the two factors are not perfectly correlated.

[8]In this particular example, γ_{35} is also unaffected ($\gamma_{35}^* = \gamma_{35}$) since $\gamma_{34} = 0$.

[9]It is shown in Chamberlain (1976) that we need at least n(n + 1)/2 restrictions for the n-factor model. Conditions on the placement of the zeros are also presented.

[10]It might appear that we could fit the variance-components

specification into the MIMIC framework by setting $\underset{\sim}{x}$ equal to a set of group indicator dummy variables. Then in the one-factor case we would set $\underset{\sim}{\beta}' = (f_1, \ldots f_q)$. This is the fixed effects version of the model. The random effects version which we use amounts to imposing on $\underset{\sim}{\beta}$ the prior distribution that the f_i are a random sample. But this approach is misleading if the sample expands by adding more groups, for then the number of parameters expands also. In fact many applications of the variance-components model have a large number of groups with only a few observations in each one. In that case using the group dummy variables to form IV's gives inconsistent estimates. For example, with

$$y_2 = \frac{d_2}{d_1} y_1 + e_2 - \frac{d_2}{d_1} e_1 ,$$

the use of group dummies in a two-stage least square calculation gives $\Sigma_i \bar{y}_{1i} \Sigma_i \bar{y}_{2i} \bar{y}_{1i} / \Sigma_i \bar{y}_{1i}^2$, where $\bar{y}_i \bar{y}_{1i}$ is the sample average over the observations in the i^{th} group. This averaging does not eliminate the bias due to the measurement error in y_1 unless there is a large number of observations in each group.

[11]The general results in Chamberlain (1976) on the identification of variance-components models also apply to the simultaneous MIMIC model.

[12]This procedure is similar to the one used by Heckman (1976) to correct for sample selection bias. See also Madalla (1976) and Lee (1976).

REFERENCES

Balestra, P. and M. Nerlove, 1966. "Pooling Cross-Section and Time-Series Data in the Estimation of a Dynamic Model: The Demand for Natural Gas," Econometrica, 34(3) pp. 585-612.

Chamberlain, G., 1976. "Identification in Variance-Components Models," Harvard Institute of Economic Research Discussion Paper No. 486.

Chamberlain, G. and Z. Griliches, 1975. "Unobservables with a Variance-Components Structure: Ability, Schooling, and the Economic Success of Brothers," International Economic Review, 16(2), pp. 422-449.

________, 1976. "More on Brothers," Harvard Institute of Economic Research Discussion Paper No. 469.

DeFinetti, B., 1937. "La Prévision: Les Lois Logiques, ses Sources Subjectives," Ann. Inst. H. Poincaré, 7(1). English translation in Kyburg and Smokler (1964).

Geraci, V.J., 1974. "Simultaneous Equation Models with Measurement Error," Ph.D. Dissertation, University of Wisconsin: Social Systems Research Institute Paper 7407.

__________, 1976. "Identification of Simultaneous Equation Models with Measurement Error," Journal of Econometrics, 4(3), pp. 263-283.

Geraci, V.J. and A.S. Goldberger, 1971. "Simultaneity and Measurement Error," University of Wisconsin: Social Systems Research Institute Paper, 7125.

Heckman, J.J., 1976. "Sample Selection Bias as a Specification Error," University of Chicago, unpublished.

Heckman, J.J. and R.J. Willis, 1975. "Estimation of a Stochastic Model of Reproduction: An Econometric Approach," in N.E. Terleckyj; ed., Household Production and Consumption. New York: Columbia University Press.

__________, 1977. "A Beta-Logistic Model for the Analysis of Sequential Labor Force Participation by Married Women," Journal of Political Economy, 85(1), pp. 27-58.

Kyburg, H.E. and H.G. Smokler, eds., 1964. Studies in Subjective Probability. New York: Wiley.

Jöreskog, K. and A.S. Goldberger, 1975. "Estimation of a Model with Multiple-Indicators and Multiple-Causes of a Single Latent Variable," Journal of the American Statistical Association, 70(351), pp. 631-639.

Lazarsfeld, P.F. and N.W. Henry, 1968. Latent Structure Analysis. Boston: Houghton Mifflin.

Lee, L., 1976. "Unionism and Wage Rates: A Simultaneous Equations Model with Qualitative and Limited Dependent Variables," University of Florida: Working Paper in Economics No. 76-77-04.

Madalla, G.S., 1971. "The Use of Variance-Components Models in Pooling Cross-Section and Time-Series Data," Econometrica, 39(2), pp. 341-358.

__________, 1976. "Self-Selectivity Problems in Econometric Models," University of Florida: Working Paper in Economics No. 76-77-06.

Mundlak, Y., 1976. "On the Pooling of Time-Series and Cross-Section Data," Harvard Institute of Economic Research Discussion Paper No. 457.

Nerlove, M., 1971. "A Note on Error-Components Models," Econometrica, 39(2), pp. 383-396.

Wallace, T.D. and A. Hussein, 1969. "The Use of Error-Components Models in Combining Cross-Section with Time-Series Data," Econometrica, 37(1), pp. 55-72.

Wiley, D.E., 1973. "The Identification Problem for Structural Equation Models with Unmeasured Variables," in A.S. Goldberger and O.D. Duncan, eds., Structural Equation Models in the Social Sciences. New York: Seminar Press.

CHAPTER 8

ON THE EFFECTS OF FAMILIES AND FAMILY STRUCTURE ON ACHIEVEMENT*

William H. Sewell and Robert M. Hauser
The University of Wisconsin-Madison

Information on the family and socioeconomic characteristics, ability, and achievement of members of the same sibship may be used to address two distinct questions: (1) How and why are siblings different from one another?, and (2) How and why are siblings more like one another than unrelated persons?

In attempting to answer the first question it is convenient to remove the effects of shared environment and heredity and to look at the influence of variables on which siblings do not have common values: birth order, birth year, and birth interval (spacing). These variables are logically related to the size of the sibship and may interact with it, so size of the completed sibship must be taken into account in an adequate research strategy. Finally, siblings may be of the same or of opposite sex, and this, too, will affect the differentiation of life-chances among family members.

In the strategy for answering the second question it is convenient to ignore the factors tending to diversify the achievements of siblings, while attempting to measure and interpret their shared background. Siblings have a partly overlapping genetic heritage. Excepting the possibility of temporal change within the family of orientation, siblings share a set of parents (and other relatives) with whom they each interact in ways which partly reflect the social and cultural divisions in the larger society. Some of these shared characteristics include education, occupation, and income of the parents, religion, ethnicity, and family size. There are other parts of the social environment, too, which do not involve the functioning of families in a narrow sense, but whose nature and influence varies from family to family. For example, the neighborhood and community in which the family resides and the schools attended by their children are of this character.

Ultimately, the division between the purposes of studying the similarity of siblings and of studying differences among them is strained and artificial. We have already noted that family size enters both analyses, as will sex. Ideally one would hope to construct a comprehensive

---

This work was supported by grants from the National Institute on Mental Health (M-06275) from the Spencer Foundation. We thank Michael Olneck and Arthur S. Goldberger for comments on an earlier draft of this paper. They are in no way responsible for its remaining faults.

model of family influences on achievement which would render the distinction unnecessary. For the moment we think the distinction is a useful heuristic device; it breaks the research problem into two parts, neither of which is especially simple when taken by itself.

Our discussion of family effects on achievement is in three parts. First, we briefly review research on the effects of families and of family structure. In this review we have deliberately avoided an elaboration of the effects of socioeconomic background variables, and we have equally neglected the methods and results of twin studies. Each of these topics is well-covered in the review paper by Leibowitz or by others of the papers in this volume. Second, we describe the content and nature of a unique body of data on family structure and the achievement of siblings which we have obtained or are presently obtaining for a large and heterogeneous panel of Wisconsin high school graduates. Last, in the context of our research design, we review issues in the analysis and interpretation of family effects on achievement when data are available for full sibships and for selected pairs of siblings.

Family Structure, Sibling Resemblance, and Achievement

At least since the time of Galton (1874), scholars have studied the effects of birth order on eminence, educational attainment, occupational achievement, aspirations and motivation, various aspects of deviance, including mental illness, delinquent behavior and alcoholism, and selected personality characteristics, such as anxiety, dependency, affiliation, achievement orientation, conformity, and measured intelligence. In the last ten years, this massive literature has been competently reviewed by a number of writers, including Sampson (1965), Altus (1966), Warren (1966), Bayer and Folger (1967), Bradley (1968), Sutton-Smith and Rosenberg (1970), Adams (1972), and Schooler (1972).

Adams has summarized a number of post-hoc theories that have been used as explanations of birth order effects, including physiological, psychological, socialization, and economic explanations. He, and many of the other writers, have pointed out that the findings to date are seriously flawed by small samples, selection bias, and failure to control for variables known to be related to sibling position and to the dependent variables under study. Moreover, none of the past studies has had adequate information to examine the influence of family structure in a sufficiently comprehensive and systematic way to permit definitive conclusions regarding the influence of sex, age, sibling position, sibship size, and spacing on career achievements.

The influences of family structure on achievement may be studied in samples of persons, as in the research of Blau and Duncan (1967), where structural variation between families is correlated with achievement variables. Also, family influences may be studied in samples of families (minimally, in at least one sib-pair from each family), as in the research of Lindert (1974), where structural variations within families may be correlated with achievement variables. The first design risks the confounding of family structural characteristics with other characteristics of the family of orientation, as in the correlations between completed family size and social class or religion. The second design implicitly controls all of the persistent characteristics of the family of orientation (whether or not we happen to know what they are), but variations in ordinal position, family size, and child-spacing are inherently confounded with temporal change in the larger society.

The best and most extensive study of between-family variations is that of Blau and Duncan (1967), which shows that both the size of parental family and the sibling position of the son exert an important influence on the son's subsequent occupational career. The attainments of first-born and last-born sons are superior to children in the intermediate positions but this advantage or disadvantage depends to some extent on family size. Sibling position and number of siblings interact in such a way that there is little difference in the achievements of oldest and youngest children in small families but youngest children are more successful than oldest in large families. Older sons in large families may make sacrifices and take on responsibilities for younger ones so that the resulting benefits accruing to younger sons compensate for the more limited resources, both psychological and economic inputs, available for any child if there are many children in the family. Almost all of the influence of family structure and climate on occupational achievement is transmitted through education. Blau and Duncan (1967: 330) conclude that "The family into which a man is born exerts a profound influence on his career, because his occupational life is conditioned by his education and his education depends to a considerable extent on his family."

Although this study is superior to any previous research on family structure and careers, both in its large and representative sample of males in the United States labor force and the sophistication with which the data were analyzed, its conclusions are limited by the fact that no data were available on the achievements of other siblings than the oldest brother and information is available only on the number of years he attended school. Moreover, nothing is known about family structure other than the size of the sibship and the sibling position of the respondent. For example, neither the 1962 Blau-Duncan survey, nor its 1973 replicate (Featherman-Hauser, 1975) contained a roster of siblings by age and sex. Further, women were excluded from the Blau-Duncan survey; they neither appear as respondents nor as members of sibling pairs.

Another important study has been reported recently by Lindert (1974), which covers a wider range of family structure characteristics, including sex, age, sibling position, family size, birth order, and spacing, for a sample of 1,087 siblings collected in 1963, by a Cornell Medical School team that interviewed 312 male employees of a New Jersey utility company in search of information about the incidence of heart disease. The respondents, aged 55-61, gave information about their siblings' age, sex, education, and most recent occupation (see Hermalin, 1969). Lindert proposes a simple explanation of the way in which family size and birth order should influence a child's subsequent attainments by governing the time and inputs the child receives from his parents (based on another sample which indicates the effect of family size on the time parents spend with young children), and tests the link between sibling position and achievement within, as well as between, families. [Zajonc and Markus,

(1975) take a related approach in their "confluence" model of the effects of family size, child-spacing, and birth order on measured intelligence.] Lindert's findings clearly support the conclusions of Blau and Duncan that sibling position and sibship size are important influences on educational and occupational achievements. He also concludes that this is probably due to the fact that sharing of family time and resources is a handicap to subsequent achievement. All family background and structure variables are found to explain schooling levels better than they explain occupational achievements.

The major weaknesses of the Lindert study are its relatively small and highly selective sample; its dependence upon respondents' reports on the achievements of their siblings; and its lack of information on such important characteristics as family background, siblings' work histories, and siblings' income. Nonetheless, the sophisticated analytic techniques and the ideas presented form a solid basis for additional research on the effects of family structure on career achievements in more representative samples, which have more complete and reliable information on socioeconomic origins, family structure, academic ability of siblings, and socioeconomic achievements.

Although the confounding of structural variables with time complicates analyses of family structure based on data from sibling rosters or sibling pairs, these data may also be used to study the similarity of siblings. That is, the resemblance of persons reared together is a fundamental indicator of the force with which the family functions to create and maintain systems of social differentiation and inequality. Sibling resemblance captures the effects of social and economic background, of family structure (to the extent it is common to all members of a sibship), and of all other commonalities of the social and psychological functioning of the family. It is possible to give sibling resemblance an explicit interpretation to the extent that shared familial characteristics have been measured. Studies of sibling resemblance have a long—and at present, somewhat dishonorable—history in connection with attempts to segregate the influences of nature and nurture. However, such data have recently been exploited effectively by students of social stratification and inequality without becoming embroiled in that issue.

Foremost among such work is that of Duncan, Featherman and Duncan (1972), who used data on the educational attainments of brothers in the 1962 OCG survey, and that of Jencks and his associates (1972), who drew on data from the 1962 OCG survey and a variety of other sources. Duncan, Featherman and Duncan showed that sibling resemblance in a variety of social background characteristics cannot explain the resemblance of brothers in the length of their schooling. [Bowles (1972) argued, we think unpersuasively, that a broader conception of socioeconomic background could account for the similarity of brothers in schooling.] Hauser and Featherman (1976) have recently replicated and elaborated this finding in the 1973 OCG data. However, neither OCG survey includes data on the occupations or earnings of brothers (or on their abilities), and this severely limits the usefulness of these surveys in studying sibling resemblance. A large 1964 NORC survey did ascertain occupations for a full (male) sibling roster, but not even educational attainments were ascertained in that study.

Jencks (1972, Appendix B) undertook an ambitious interpretation of sibling resemblance data for males in the U.S., which attempted to model occupational and economic as well as educational achievements. In many respects this analysis serves as a methodological template for future research, but Jencks' analyses highlight the paucity of complete or representative sibling resemblance data. He had to draw on a variety of source materials, many of which are not strictly comparable. Jencks' analyses highlight the data needs for studies of sibling resemblance. Minimally, measures are needed on social background, cognitive ability, educational attainment, occupational achievement, and earnings for sibling pairs drawn from a single representative sample. Michael Olneck's (1976) study of brothers in Kalamazoo, Michigan, is an exemplary effort to bring together the required data in a local sample.

Background of the Wisconsin Study

Before elaborating the design of our research on siblings, a brief review of our study of the influence of social and psychological factors on the educational, occupational and economic careers of the men and women in our sample is in order. The basic data with which the project began were obtained from a statewide survey of the post-high school plans of all seniors in public, parochial, and private high schools in Wisconsin in 1957. The questionnaire contained information on a number of aspects of (a) the student's social background, including the occupation, education, financial status, and place of residence of his/her parents, and the extent to which parents encouraged education beyond high school; and (b) the student's educational background, including the pattern of high school courses, the educational plans of friends, general attitudes toward education, and educational and occupational plans. A random sample of 10,317 cases was drawn from the 1957 survey and data from school records, the census, and other public sources were added to the files of each student. The latter included measured intelligence, high school rank, and a four-year average of parental income.

These data provided the basis for a great deal of work on the effects of such social background variables as sex, socioeconomic status, community origins, and neighborhood contexts, and of such social psychological variables as measured intelligence, rank in high school class, and the student's perception of the expectations of significant others (parents, teachers, and peers) on the educational and occupational aspirations of youth. This analysis led to several papers which examine the effects of the variables in question on educational and occupational aspirations, usually controlling for sex, socioeconomic status, and measured ability (Sewell, 1964; Sewell and Orenstein, 1965; Sewell and Armer, 1966; Sewell and Shah, 1968).

One of the original aims of the Wisconsin study was to determine not only the effects of social background factors on aspirations but also on actual educational, occupational, and economic attainments. With this end in view, we launched a follow-up study in 1964, seven full years after the students in our sample graduated from high school. In this survey we sought information from the parents of the sample members on the educational histories, occupational attainments, military service, marital status, and current geographical location of each student. We obtained data for 87.3 percent of the members of the sample. We later obtained the earnings histories of the males in our sample from the Social Security Administration, using linkage techniques guaranteeing the anonymity of the information. A number of publications addressing central issues in social stratification have resulted from our analyses of these data.

We were especially interested in learning more about how social origins affect educational attainment particularly because of the crucial role that education plays in all later attainments. The first major study we did after obtaining the follow-up data involved a thorough examination of the effects of socioeconomic origins on the attainment of higher education (Sewell and Shah, 1967). We assessed the influence of family socioeconomic status controlling for measured intelligence, for males and females separately, as they progressed through the process of higher education: from college plans, to college attendance, to college graduation. Our results indicate that, for each sex, socioeconomic status is an important determinant at each level of educational attainment, even when intelligence is controlled. For both sexes intelligence becomes more important than socioeconomic status as progress is made through the educational system, but at no point does one's socioeconomic status cease to be an important determinant of the attainment of the next step in the process.

During the past several years the major thrust of the analytic work on the project has been directed toward the development of causal models of the status attainment process. In our earlier work we had identified a number of experiences that young people undergo in their formative years which have an important bearing on post-high school educational outcomes. These include level of performance in high school, whether significant others encourage or discourage high educational and occupational aspirations and whether the students actually develop these aspirations. All of these experiences are affected by the social origins, academic ability, and sex characteristics of the individual and become the mechanism through which these background characteristics transmit their influence. In addition, these same social psychological experiences have direct and indirect effects of their own, quite independent of the person's background characteristics.

This complex process has been the focus of much of our recent research, and we have been developing and testing structural equation models to further explicate the process of attainment. Building on the work of Blau and Duncan (1967), we have devised and published a linear recursive model that attempts to elaborate and explain the effects of socioeconomic origins and academic ability on educational achievements and occupational attainments as these influences are mediated by social psychological processes (Sewell, Haller and Portes, 1969; Sewell, Haller and Ohlendorf, 1970). This model links socioeconomic status and academic ability with educational and occupational attainment by means of the social psychological variables previously mentioned. The model demonstrates that socioeconomic status has no effect on performance in high school independent of academic ability, but has strong direct and indirect effects on significant others' influence, educational and occupational aspirations, and through these on educational and occupational attainments. The role of academic ability is somewhat different in that it has strong direct effects on high school performance, independent of socioeconomic status, and direct and indirect effects on significant others and on educational and occupational attainments. This model succeeds in explaining 57 percent of the variance in post-high school educational attainment and 40 percent of the variance in early occupational attainment for the boys in our sample.

Recently we have further elaborated and extended our model by disaggregating socioeconomic status into its component parts--parents' income, mother's education, father's education, and father's occupation—and by decomposition of significant others' influence into parental encouragement, teachers' encouragement, and peers' plans and by adding son's earnings in 1967 as the final dependent variable. This enables us to obtain estimates of the individual role of each of the independent variables in the educational and occupational, and economic attainment process. Because this analysis is quite complicated, we shall not present our detailed findings in this review. A complete summary and discussion of our findings is given in our recent book (Sewell and Hauser, 1975). Perhaps it is sufficient to say that although we were very pleased with the power of our model not only to explain educational and occupational achievement in the early career but also to interpret the influences of family background in the achievement process, we were quite disappointed with its lack of power in predicting earnings: It explains less than 10 percent of the variance in the 1967 earnings of employed males.

We believe that its poor showing in explaining earnings is in part due to the fact that our data pertain to an early period in the earnings careers of our sample. For those who have gone on to college and professional schools, it will take several years for their earnings to reflect their high levels of education (see Hauser and Daymont, 1976). Another factor is that we need more proximate and pertinent information about career experiences, including current jobs, labor market areas, on-the-job training, graduate and adult education, family formation, and military service, if we are to explain earnings more adequately.

To obtain this and other necessary information on the careers of our sample, at approximately age 35, we carried out an extensive follow-up of a 45-minute telephone interview. The interview schedule was modeled after that used in the 1973 National Mobility Survey by Featherman and Hauser (1975), but it is longer and more intensive and it includes women as well as men regardless of marital status. It covers regular and non-regular schooling histories; selected occupation reports from first job through current job; military service and training; current labor force status, work experience, and earnings; job characteristics, satisfactions and aspirations; social participation; marital and fertility experience; women's job histories; aspirations for children; sibling characteristics (see below); and retrospective reports and addenda to our earlier information on social background, aspirations, and peer influence.

A tracing operation carried out in 1974 successfully located 97.4

percent of the original members of the sample; this figure included 99 percent of the 9007 persons for whom responses were obtained in the 1964 survey and 86.2 percent of the 1310 persons for whom no responses had been obtained in the 1964 survey. Ultimately, 88.6 percent of the members of the original sample were interviewed by telephone in 1975. This compares favorably to the 87.3 percent of the original 10,317 people for whom data were obtained in 1964. Those interviewed in 1975 comprised 91.4 percent of the original members of the sample who were not known to be deceased, disabled, or currently living outside the United States. The response rate is slightly higher for females than males, 92.3 percent and 90.5 percent, respectively. The response rate is 93.5 percent among persons whose parents responded in 1964 and 77.0 percent among 1964 nonrespondents; this differential is primarily a result of tracing failure. [For further details of the tracing and interviewing processes, see Clarridge, Sheehy, and Hauser (1976).]

Wisconsin Sibling Data

In the 1975 survey men and women were asked the number of older and younger sibs of each sex. For all living bothers and sisters, age, sex, first name, and educational attainment were ascertained, and one sibling was designated to be the subject of additional queries. The designated sibling was always the respondent's twin, if he or she were a twin. Otherwise, a sibling was designated at random from a selection table pasted into the interview. For the designated sibling we also ascertained the name and location of the last high school attended (if any), labor force status, occupation, industry, class of worker, and current full name and address. At least one living sibling was reported by 92 percent of respondents; 7 percent of respondents were only children, and the remainder were nonrespondents or sole survivors of their sibships. Item response rates were high for the educational and occupational characteristics of siblings, and 96 percent of the designated siblings were reported to have attended a Wisconsin high school. Of respondents with living siblings 90 percent reported at least a city or town of current residence of the designated sibling; because of our access to parents and other informants we do not think this necessarily bounds our ability to locate named siblings.

Assuming that the proxy data are sound, we can analyze sibling similarity and effects of sex, sib position, and sib composition on schooling of the full sibship. For the designated sibs we can look at labor force participation and occupation as well as schooling, and we can bring measured intelligence into the picture to the extent that we can locate test scores (see below). With these data we cannot treat earnings or income as a dependent variable, nor any other social or psychological characteristic (except, perhaps, geographic proximity).

In addition to obtaining the proxy data for full sibships and designated siblings we are presently designing a survey of a subsample of about 2000 designated siblings for which the fieldwork will be carried out early in 1977. Our specification of the composition of sibling pairs to be included in this survey reflects a balance between our interest in assessing the quality of proxy reports about siblings and in obtaining data on sibling pairs which has value independent of the 1975 panel survey. For example, we need some opposite-sex sibling pairs in order to assess the effects of sex on proxy reports of siblings' characteristics. Also, from opposite-sex pairs we can assess the sex-specificity of global family influences. We know that families do induce substantial resemblance in at least the educational achievement of brothers. There is less evidence about women and their sisters and women and their brothers, and we suspect there may be less resemblance as well. If so, measures of filial resemblance, taken alone, may overstate the conservative force of the family in American society. At the same time opposite-sex pairs are less likely than male-male pairs to yield paired data on current occupations (because many women are not in the labor force), nor will opposite-sex pairs give us the detailed data on family formation processes which we can obtain from female-female pairs. We plan to obtain interviews with 750 same-sex designated sibs of each sex (total of 1500 interviews) and 250 opposite-sex designated sibs of each sex (total of 500 interviews). In addition we will interview all twins of sample members who were not interviewed in the 1975 survey or selected into one of the subsamples of designated siblings. Within the four strata defined by sex of the respondent and of the designated sibling, the sample is stratified by size of the sibship and by ordinal position and educational attainment of the designated sibling. Thus, we expect our sampling design to be considerably more efficient than a simple random sample.

In the 1977 survey of designated siblings, we will measure educational attainment in detail; additional information about secondary schooling will help us in locating test scores. We will ascertain current labor force status, occupation and other job characteristics, work experience and earnings in 1976, and other variables drawn from the 1975 instrument, including military service and vocational training, first and 1970 occupations, marital and fertility histories, aspirations for themselves and their children, and levels of social participation. In addition to these replicate items we shall ascertain whether the designated siblings were full, step, or adoptive siblings of our 1975 respondents and whether the sibling pair were reared in the same household.

Because of its possible importance as a component of sibling resemblance, we are obtaining the designated sibling's Henmon-Nelson score, or other mental ability test score (which we already have for members of the original samples). These are accessible through the Wisconsin State Testing Program, which is affiliated with the University of Wisconsin. Test scores are stored on campus in bound volumes for each semester and by locality, school, and grade-level within volume. Listings include identification of the test on which the score was obtained. In virtually every case test scores on the full sample were obtained from the Henmon-Nelson test administered in the 11th grade, but in more recent years (than 1956, the full sample's junior year) coverage has declined, and we may have to use scores on other tests. Also, because of variations in test coverage (and because some siblings did not reach the junior year in high school) in some cases we shall ascertain test scores at other grade levels.

We have carried out a very preliminary pilot effort to ascertain test scores for a sample of over 200 designated siblings about whom information was reported by our respondents, using only the files of the Wisconsin State Testing Service, and we have found, without exhaustive effort, test scores for almost two-thirds of those in the sample. We believe that we can improve considerably on this record by using additional information from our interviews with the designated siblings. We will also be able to organize and systematize our search procedures more effectively once we get into production. We do not think it over-optimistic to expect to obtain test scores for 75 to 80 percent of the designated siblings.

Effects of Families and Family Structure

The structure of the completed or proposed Wisconsin data is complex, and the potentials and problems in analyzing the data depend on which persons and variables are involved. In the completed, cross-sectional survey of Wisconsin high school graduates we have a complete roster of the living siblings of each 1975 respondent. We know the labor force status and occupation of one randomly designated sibling of each respondent, and we propose to find an ability measure for each designated sibling as well. In the subsample of designated siblings we will have self-reports of education and occupation, as well as other variables drawn from the instrument used in the original sample. Thus, there is an inverse relationship between the size or scope of the sample—full sibships, all respondent-random sibling pairs, and a subsample of respondent-sibling pairs--and the scope of the data available for both members of the pair.

What are the main analytic problems in the analysis of family or sibling-pair data? The first problem is to describe the data in a way that mirrors the effects of family structure and family membership in a full sibship of selected families. Second, we discuss the use of data on selected pairs of sibs, rather than full sibships, and the selection of a sample of families in which at least one sibling completed 12 years of school. Finally, we consider the inclusion of other variables in the analysis, which may be additional factors in a single-equation model or intervening variables in a multi-equation causal model.

In brief, how does one analyze sibling or family data? The effects of family structural variables may be measured using statistical models for the analysis of variance or (equivalently) dummy variable regression analysis. Variables like size of the sibship, birth order, sex, birth year, spacing between births and prior achievements enter the models as explanatory factors or regressors. Most importantly, the main effect of family membership enters the design as a separate factor in the analysis of variance or as an additional regressor in the dummy variable regression analysis. It is the use of family membership as a factor in the design which distinguishes our analysis from those which might be carried out in a cross-section of persons. Whether we use data on sibling pairs or full sibling rosters, this design eliminates confounding of family size and birth order with other characteristics of families. These analyses might be carried out in other ways, e.g., by taking deviations of individual outcomes from family averages (see below), by taking differences between outcomes in selected sibling pairs, or by taking differences between weighted averages of outcomes in different birth orders. All of these schemes will produce equivalent results, if used correctly. However, we believe the analysis of variance design is easiest to communicate, while a regression set-up is the most convenient and efficient in practice.

The effects of family membership may be analyzed by measuring the resemblance of siblings. A gross measure of intra-family resemblance may be obtained by averaging differences between siblings in respect to an outcome measure. The variability in average deviations between siblings will differ from that in average deviations between non-siblings as a mathematical function of the correlation between siblings on the outcome in question. That correlation reflects a balance between the factors which make siblings alike and different from one another. Sibling correlations may be analyzed in detail using structural equation models which explicitly or implicitly represent the processes which make siblings resemble each other or differ from one another. Thus, the statistical analysis of sibling correlations ultimately explains and interprets the observed differences between siblings (see equation 7 below).

Without attempting to exhaust either the problems or potentialities in the Wisconsin data we shall describe some possible analyses of family and sibling-pair data. This description is not very refined or rigorous, and we intend it merely to indicate the feasibility and the style of the analyses.

Size and Ordinal Position in Full Sibships

We begin with a very simple model of the effects of birth order and family size on a single outcome variable in a full sibship. Consider a single outcome, X, e.g., educational attainment, whose value may be denoted by X_{ijk} for the member of the i^{th} family in the j^{th} birth order in families of sibship size k. Assume that in each of k family sizes, $k>1$, we have observed all of the X_{ijk} in a sample of families. That is, the situation resembles that of our full roster of siblings' educational attainments. For the moment we ignore the effects of all variables not explicit in the design, e.g., sex, birth year, social background, or whether the observation pertains to a respondent or to one of his or her siblings. Families are nested within sibship sizes, and persons within families. Within each sibship size the design is a complete crossed and balanced two-way layout with one observation per cell. However, because $0< j\leq k$ for all k, the overall design looks like a staircase. There is variation in the number of sibships of each size and, depending on the latter, in the number of persons in each birth order, even though there is a uniform distribution of birth orders within any family size. We may write the model as

$$X_{ijk}=\mu + a_i + \beta_j + \gamma_k + \delta_{jk} + e_{ijk} \qquad (1)$$

where μ = grand mean of X

a_i = effect of the i^{th} family, a random variable with $E(a_i) = 0$

β_j = effect of the j^{th} birth order, $\sum_j n_j \beta_j = 0$

γ_k = effect of the k^{th} sibship size, $\sum_k n_k \gamma_k = 0$

δ_{jk} = effect of the jk^{th} birth order-family size combination,

$$\sum_j n_j \delta_{jk} = \sum_k n_k \delta_{jk} = 0$$

e_{ijk} = a random disturbance, $E(e_{ijk}) = 0$

Note that this is a mixed model, since birth order and sibship size have fixed effects, while family is a random variable. We include possible interactions between birth order and sibship size, but nesting of families within sibship sizes precludes an ik interaction, and with one observation per cell the ij interactions within levels of k are assumed to reflect error.

To avoid the normalization problems imposed by the staircase design, we specify the model for fixed k as

$$X_{ijk} = \mu_k + a_i + \beta_{jk} + e_{ijk}, \qquad (2)$$

so $\mu_k = \mu + \gamma_k$, $\beta_{jk} = \beta_j + \delta_{jk}$, and $E(a_i) = E(e_{ijk}) = \sum_j^k \beta_{jk} = 0$

Further, let $\bar{X}_{.jk}$ = mean achievement in the j^{th} birth order = $\sum_i^{n_k} X_{ijk}/n_k$

and $\bar{X}_{..k}$ = mean achievement in sibships of size k = $\sum_i^{n_k} \sum_j^k X_{ijk}/kn_k = \sum_j^k \bar{X}_{.jk}/k$.

We may estimate β_{jk} as

$$\hat{\beta}_{jk} = \bar{X}_{.jk} - \bar{X}_{..k} = (\mu_k + \sum_i^{n_k} a_i/n_k + \beta_{jk} + \sum_i^{n_k} e_{ijk}/n_k) \qquad (3)$$

$$- (\mu_k + \sum_j^k \sum_i^{n_k} a_i/kn_k + \sum_j^k \beta_{jk}/k + \sum_j^k \sum_i^{n_k} e_{ijk}/kn_k)$$

$$= \beta_{jk} + \sum_i^{n_k} e_{ijk}/n_k - \sum_j^k \sum_i^{n_k} e_{ijk}/kn_k,$$

since $\sum_j^k \beta_{jk} = 0$, and $\sum_j^k \sum_i^{n_k} a_i/kn_k = k\sum_i^{n_k} a_i/kn_k = \sum_i^{n_k} a_i/n_k$. That is, even though the family effect is a random variable, it does not bias, nor contribute to variance in our estimator, $\hat{\beta}_{jk}$, of the birth order effect.

Finally, $E(\hat{\beta}_{jk}) = \beta_{jk}$; that is the difference between mean achievement in a birth order and the mean for all birth orders is an unbiased estimator of the effect of the birth order.

If the model of equations 1 and 2 were correct, with minor modifications it could be applied to outcomes measured only for sibling pairs as well as to the full sibling rosters in the Wisconsin data (for which we have measured educational attainment). That is, if we had first sampled sibships and then random pairs of persons within sibships, we would only have to deal with a problem of randomly deleted measurements of achievement. Of course we always know the structural characteristics of the full sibship--which may be obtained from birth order, sex, and birth year.

However, as we argue below, the model of equations 1 and 2 cannot be applied to the Wisconsin data (for full sibships or for respondent-sibling pairs) because respondents will not be uniformly distributed across birth orders within sibships of any size and because there is a lower bound on the schooling of respondents, but not on the schooling of other members of their sibships. It is convenient to pursue these and other analytic problems in a regression framework rather than in the notation of models 1 and 2.

Practicalities and Elaborations

As a practical matter there is no need to treat each family as an explicit level in an analysis of variance design. There are too many families to make this convenient. It is similarly impractical to enter a dummy variable for each family in a regression equation which also includes other explanatory variables. However, exactly the same result may be achieved by the simple procedure of regressing individual deviations in achievement about the sibship mean on deviations in each regressor about the sibship mean (see Hauser, 1971; and Hauser, Sewell and Alwin, 1976 for applications of this idea in the measurement of school effectiveness). That is, the within-family regressions will give exactly the same effects of birth order and its interactions with family size as will the models of equations 1 and 2. The same procedure may be applied to data for individuals in full sibships or in respondent-sibling pairs. Of course the degrees of freedom in the standard regression output will have to be adjusted to reflect the fact that we will have used a degree of freedom in computing each family mean.

The effects of sibship size may be modeled in an auxiliary regression, to which we may add other variables affecting both family size and mean family achievement. For example, parental socioeconomic standing, ethnicity, or religion may account for part or all of the association between family achievement levels and sibship size. This part of the analysis can be carried out as a separate operation because these other explanatory variables do not vary from sibling to sibling within the same family. It may be desirable to estimate adjusted family achievement levels at some point in the analysis, where the adjustments are for differences in family composition on variables which vary both within and

We have thus far assumed that siblings have no personal history beyond a set of initial conditions--family size, birth order, and sex. Of course, this is not the case. We are concerned with multiple outcome variables, and we assume that outcomes later in the socioeconomic life cycle are conditional on earlier achievement. That is, we want to measure the total effects of family structural variables on achievements and, also, the direct effects, above and beyond those by way of earlier achievements. (See Alwin and Hauser,1975, for an exposition of this nomenclature of effects.) For example, as a tentative hypothesis we might expect that the effects of family structural variables on occupational status are entirely attributable to their effects on the length of schooling. In this case the analytic procedure is straightforward. We simply add the earlier outcome of each sibling to the equation for the later outcome and see what happens to the coefficients of family structural variables. If the latter coefficients then take on negligible values, the working hypothesis is sustained.

Excepting the rarity of multiple births, biological sibships do not come ready-made, but are established over a period of years. For the most part variation in birth order net of temporal change is not only counterfactual, but could not be otherwise. Thus, variation in achievement within families is partly attributable to secular changes in the larger society. Among births at a given parity in a given year the birth order of later-born siblings is confounded with year of birth and with spacing between births, and year of birth and spacing are indistinguishable. Moreover, year of birth and respondent status are obviously closely related in the Wisconsin sample. Fortunately, we believe there is enough independent variability in each of these variables within the Wisconsin sample to permit estimation of their separate effects.

Obviously, the most troublesome case is the analysis of sibling pairs of which one member is always a respondent. However, there is variability in year of birth among respondents. Roughly 20 percent of respondents were born in 1938 or earlier and 4 percent were born in 1940 and later. Moreover, in any given comparison of a respondent and a sibling of given birth orders, the respondent is the earlier-born in some cases and the later-born in others, and spacing varies within (and across) birth-order contrasts.

Birth-year effects may be ascertained by entering a dummy variable for each year of birth or, perhaps, for groups of adjacent years. Here, as in the case of respondent status, the composition of sibships by birth year affects mean achievement levels of sibships, and a post-hoc adjustment of achievements in sibships of different sizes may prove interesting. In general, larger sibships have been born over a wider span of years and, perhaps, earlier or later than smaller sibships. As a first approximation to the effects of spacing we may enter measures of the interval to the next higher and to the next lower order birth within birth orders where there are adjacent higher- or lower-order births. That is, we assume that year and spacing effects are additive across birth orders. In runs across sibships of varying sizes we may have to introduce a separate dummy variable for last-borns in order to avoid averaging in the non-effects of interval to the (non-existent) next higher order birth. This is unnecessary between families, like birth-year and spacing.

Once family has been eliminated as an explicit factor in the design, it becomes easier to look at the effects of other variables. Several variables might be added to the design, e.g., a respondent designator, sex, prior achievement, or birth-year. Throughout the following discussion it should be understood that we are always referring to within-family regressions, even where that is not stated explicitly.

By construction, at least one member of each sibship in the full sibling roster must have been a senior in a Wisconsin high school in 1957. Obviously, this affects the characteristics of families in the Wisconsin sample as compared to other possible samples of familes. For example, one might expect the Wisconsin sample of families to be better off on the average, say, than the families of a sample of sixth graders in Wisconsin schools in 1951. Moreover, within the Wisconsin family sample the achievement of respondents in the 1957 sample might differ from those of their siblings. Respondents, but not their siblings, must have reached the end of the senior year of high school. If respondents are selected on achievement, and respondents are not randomly distributed over birth orders for sibships of each size, the effects of birth order will be distorted.

In fact there are demographic reasons why the distribution of respondents over birth order may not be uniform within any given family size. That is, families were classified by sibship size after the fact, while the occurrence of a birth in a given year is a function of family characteristics and general social and economic conditions, including the distribution of potential mothers by parity. For example, in 1947, 43 percent of births were first-born, so a sample of persons in that cohort would include a disproportionate number of first-born respondents (Schooler, 1972). Moreover, even if births in 1938-39 (when most of the Wisconsin respondents were born) were uniformly distributed by parity within completed sibships, we would expect the distribution of survivors to high school graduation in 1957 to be different if there are any effects of birth order on achievement. That is, because there is a lower bound on the schooling of those in the original sample, respondents should be represented disproportionately in birth orders favorable to high levels of schooling.

With these possibilities in mind we will add a dummy variable to distinguish respondents in the 1957 survey from other siblings. The respondent effect may vary from one sibship size to another. If the effect is substantial, we shall carry out a post hoc adjustment of mean achievement levels at each sibship size. We take some comfort in the fact that selection of the respondents is explicit on education. That is, respondents should differ from the larger pool of siblings only because of the floor on their educational attainment. In analyzing achievements subsequent to schooling there should be no effect of respondent status once schooling is entered as a regressor. There is no problem in representing the effects of sex in the within-family regressions. We simply add a variable for each individual which is the deviation of a dummy variable for his/her sex from the proportion of one sex in the sibship.

in the case of first-borns, for whom there will already be a birth-order dummy variable, regardless of completed sibship size.

Sibling Resemblance

In studying the effects of family structural variables we have attempted to free our observations of the main or common effect of family membership. The object in studying sibling resemblance is to measure and interpret the main effect of family membership. From existing theories and data we would expect that commonalities of measured socioeconomic background and ability will not fully account for the resemblance of siblings in respect to achievements like schooling, occupational standing, or earnings. That is, families have coherent and persistent patterns of interaction and organization which influence the life-chances of their members and which do not merely reflect their position in the hierarchies of social or intellectual standing. With data like those we have collected or will collect for the Wisconsin sample, we can specify the relative importance of observable and not directly-observable family characteristics, and we can measure the relationships between observable and unobservable characteristics. In this way we may obtain some clues about the content or meaning of the unobserved family characteristics.

In this exposition we shall, for the most part, ignore the possible influence of family structural variables which tend to induce variation in the achievement of siblings. However, such variables may be brought into the analysis at some points to help us to specify models in which the effects of unobserved family characteristics can be estimated from empirical data. Moreover, it should be kept in mind that structural and temporal differentiation within families (as well as genetic differentiation, including that by sex) place an upper limit on the degree to which siblings can resemble one another.

How does one measure and interpret sibling resemblance? Let x_{Ri} be the achievement, say, the educational attainment, of the i^{th} respondent, and let x_{Si} be the achievement of a randomly selected sibling. As a gross measure of sibling resemblance, we might take the differences between siblings, $x_{Ri} - x_{Si}$ over all respondents (families) in the sample, say, $\bar{x}_{Ri} - \bar{x}_{Si}$. How do we know whether this average difference indicates a lot of resemblance among siblings? By definition,

$$\begin{aligned} \mathrm{Var}(x_{Ri} - x_{Si}) &= \mathrm{Var}(x_{Ri}) + \mathrm{Var}(x_{Si}) - 2\mathrm{Cov}(x_{Ri}, x_{Si}) \\ &= \mathrm{Var}(x_{Ri}) + \mathrm{Var}(x_{Si}) - 2r_{RS}\, \mathrm{Var}(x_{Ri})\, \mathrm{Var}(x_{Si}), \end{aligned} \quad (4)$$

where $\mathrm{Var}(x_{Ri} - x_{Si})$ = the variance of the inter-sibling difference in achievement,

$\mathrm{Var}(x_{Ri})$, $\mathrm{Var}(x_{Si})$ = variances of respondent's and sibling's achievements, respectively,

$Cov(x_{Ri}, x_{Si})$ = covariance of the two siblings' achievements,

and r_{RS} = Pearsonian correlation between the achievements of the two siblings. Variances are inherently positive quantities, and we expect the covariance (and correlation) between siblings' characteristics to be positive as well. If siblings' achievements were uncorrelated, the variance of intra-pair differences would be just

$$Var(x_{Ri} - x_{Si}) = Var(x_{Ri}) + Var(x_{Si}), \qquad (5)$$

and if both respondents and siblings were drawn randomly from full sibships, we would have $Var(x_{Ri}) = Var(x_{Si})$, so equation 5 would become $Var(x_{Ri} - x_{Si}) = 2Var(x_{Ri}) = 2Var(x_{Si})$. From equation 4 it is obvious that the variance of the intra-pair difference decreases as the correlation between siblings' achievements takes on larger positive values. In the limiting case where the correlation is unity, the variance of the intra-pair differences falls to zero. Thus, the variance of inter-sibling achievement differences is an equivocal measure of sibling resemblance, except where it can be compared with the sum of the variances in the achievements of each sibling. In turn the outcome of that comparison depends on the correlation between siblings' achievements. For this reason our analysis will focus on the measurement and interpretation of correlations between the achievements of siblings.

It is convenient to describe the analysis of sibling correlations using the method of path coefficients (even though we will ultimately depart from several of the conventions of that method in the estimation of effects and the presentation of results). Consider the simple path diagram in Figure 1, which represents the following structural equation model:

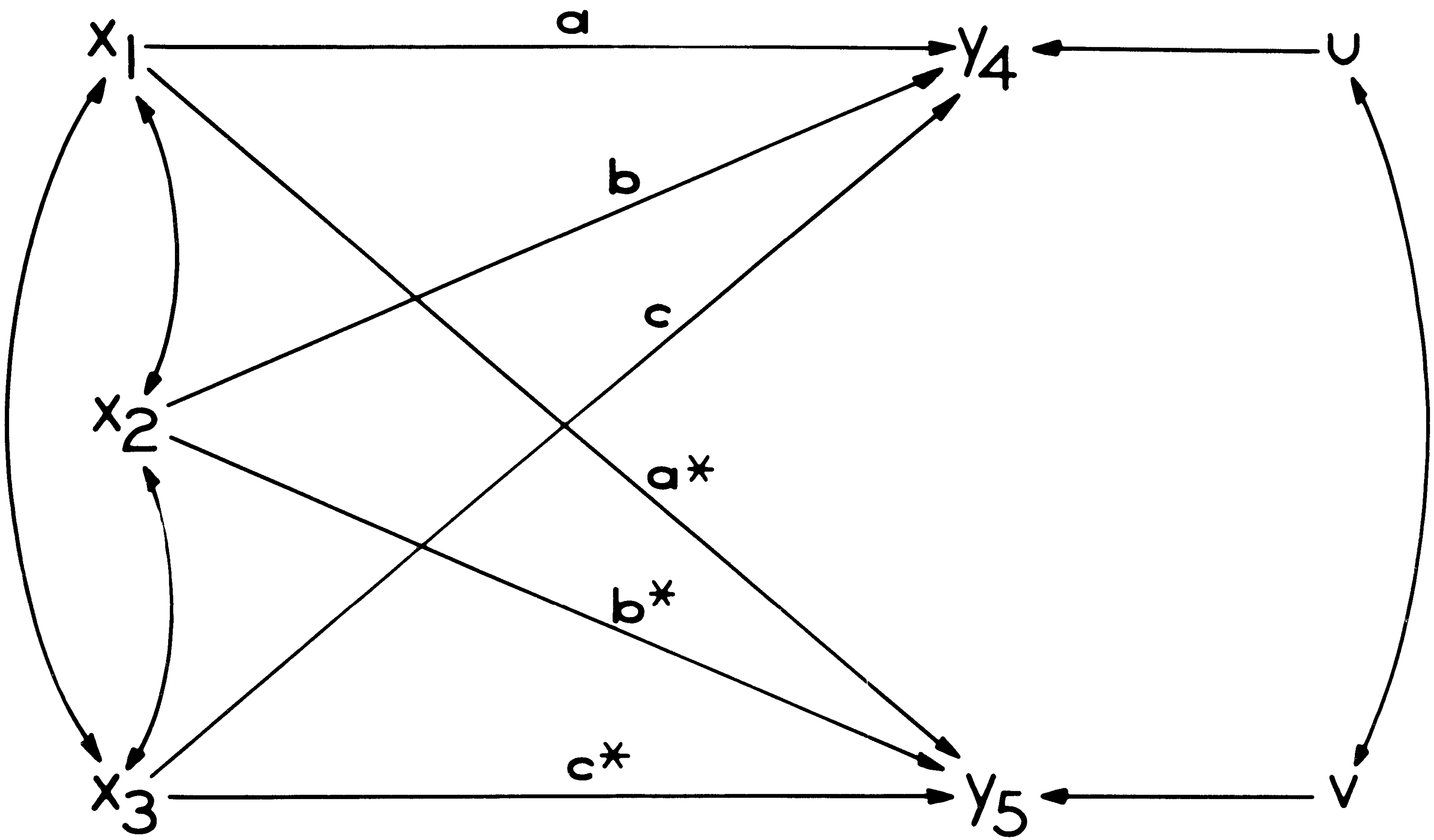
$$Y_4 = ax_1 + bx_2 + cx_3 + p_{4u}u$$

and $$Y_5 = a^*x_1 + b^*x_2 + c^*x_3 + p_{5v}v, \qquad (6)$$

where Y_4 and Y_5 are measures of the same achievement variable, say, years of schooling, for members of a sibling pair, and x_1, x_2, and x_3 are measures of family background characteristics shared by the sibling pair, e.g., mother's education, father's occupational status, and family income. For convenience of exposition the variables are understood to be expressed in standard form (as deviations from sample means, divided by sample standard deviations), but this is not intrinsic to the method.

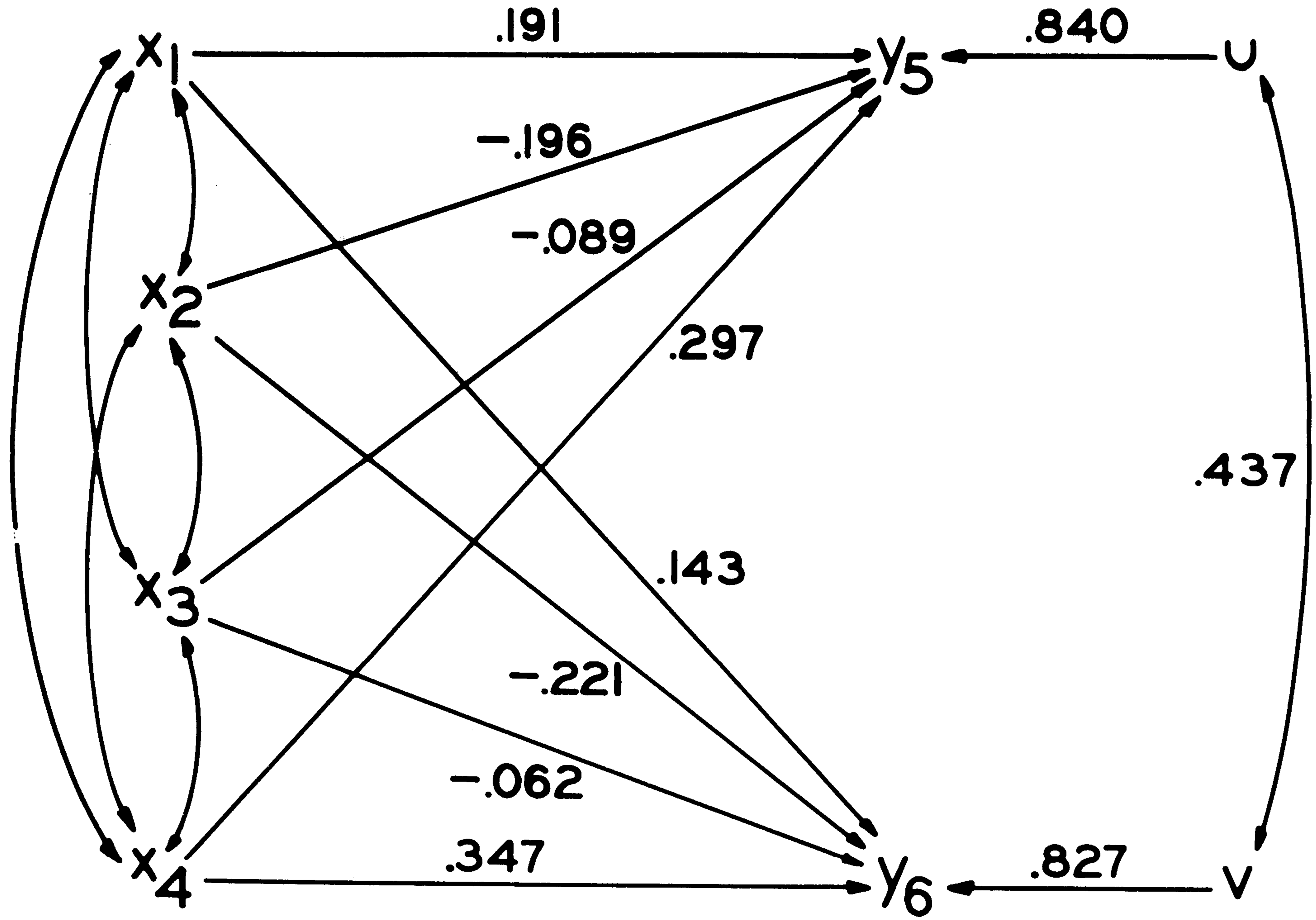
As indicated by the straight, unidirectional arrows, the achievement of each sibling is taken to depend on all 3 background variables, and in addition some variability in the achievement of each sibling may be attributable to random disturbances (u,v), whose meaning we cannot specify directly, but which are taken to be uncorrelated with the background variables. As indicated by the curved, bi-directional arrows connecting x_1, x_2, and x_3, there are presumably correlations among the background variables, but these are taken as observed and not further

Figure 1: A path model of the effects of measured family background variables on the achievement of siblings



NOTE: x_1, x_2, x_3 are measured family background characteristics of both siblings, and y_4 and y_5 are achievements of the respondent and a randomly selected sibling, respectively.

Figure 2: Effects of family background variables on the schooling of OCG respondents and their best living brothers: U.S. men aged 35 to 39 in March 1973 (N 1800)



NOTE: x_1 = father's occupational status, x_2 = number of siblings, x_3 = broken family, x_4 = father's schooling, y_5 and y_6 = schooling of OCG respondent and oldest brother, respectively.

interpreted in the model. There is a symmetry of structure between the top and bottom halves of the diagram, but we need not assume equality in the coefficients of the background variables between the two siblings. Thus, in general $a \neq a^*$, $b \neq b^*$, $c \neq c^*$. However, if sibling pairs were drawn at random from a population of families, one might wish to impose a restriction of equality between coefficients in the upper and lower portions of the diagram. We shall see below that the assumption of equality (or some other determinate relationship) between certain coefficients of variables affecting each member of a sibling pair must be invoked in more complex models of sibling resemblance.

Finally, it is evident from the diagram that u and v may be correlated. That is, whatever affects the achievement of persons beyond measured background variables, it may be the same thing, or at least partly the same thing, for members of the same sibship. The assumption that the correlation of disturbances (r_{uv}) reflects the functioning of unmeasured family background characteristics is basic to the specification of more complex models of sibling resemblance.

By repeated application of the basic theorem of path analysis we may write the correlation between the observed y-values of siblings as

$$\begin{aligned} r_{45} &= ar_{15} + br_{25} + cr_{35} + p_{4u}r_{uv}p_{5v} \\ &= a(a^* + b^*r_{21} + c^*r_{31}) \\ &+ b(a^*r_{21} + b^* + c^*r_{32}) \\ &+ c(a^*r_{13} + b^*r_{23} + c^*) \\ &+ p_{4u}r_{uv}p_{5v} \\ &= [aa^* + bb^* + cc^*] \\ &+ [(ab^* + ba^*)r_{12} + (ac^* + ca^*)r_{13} + (bc^* + cb^*)r_{23}] \\ &+ [p_{4u}r_{uv}p_{5v}] . \end{aligned} \tag{7}$$

The term in the first pair of brackets gives the component of the sibling correlation attributable to the direct effects of common causes. The term in the second pair of brackets involves the effects of correlated common causes, and the last term gives the contribution of the disturbance correlation, i.e., unmeasured shared aspects of family background.

In general we cannot expect the effect of unobserved family background variables to be negligible. For example, Figure 2 shows regressions of the schooling of 35-to-39-year-old U.S. men and their oldest (not necessarily older) brothers on four socioeconomic background variables (father's occupational status, number of siblings, broken family, father's schooling). The data pertain to about 1800 respondents in the 1973 Occupational Changes in a Generation survey carried out by the U.S.

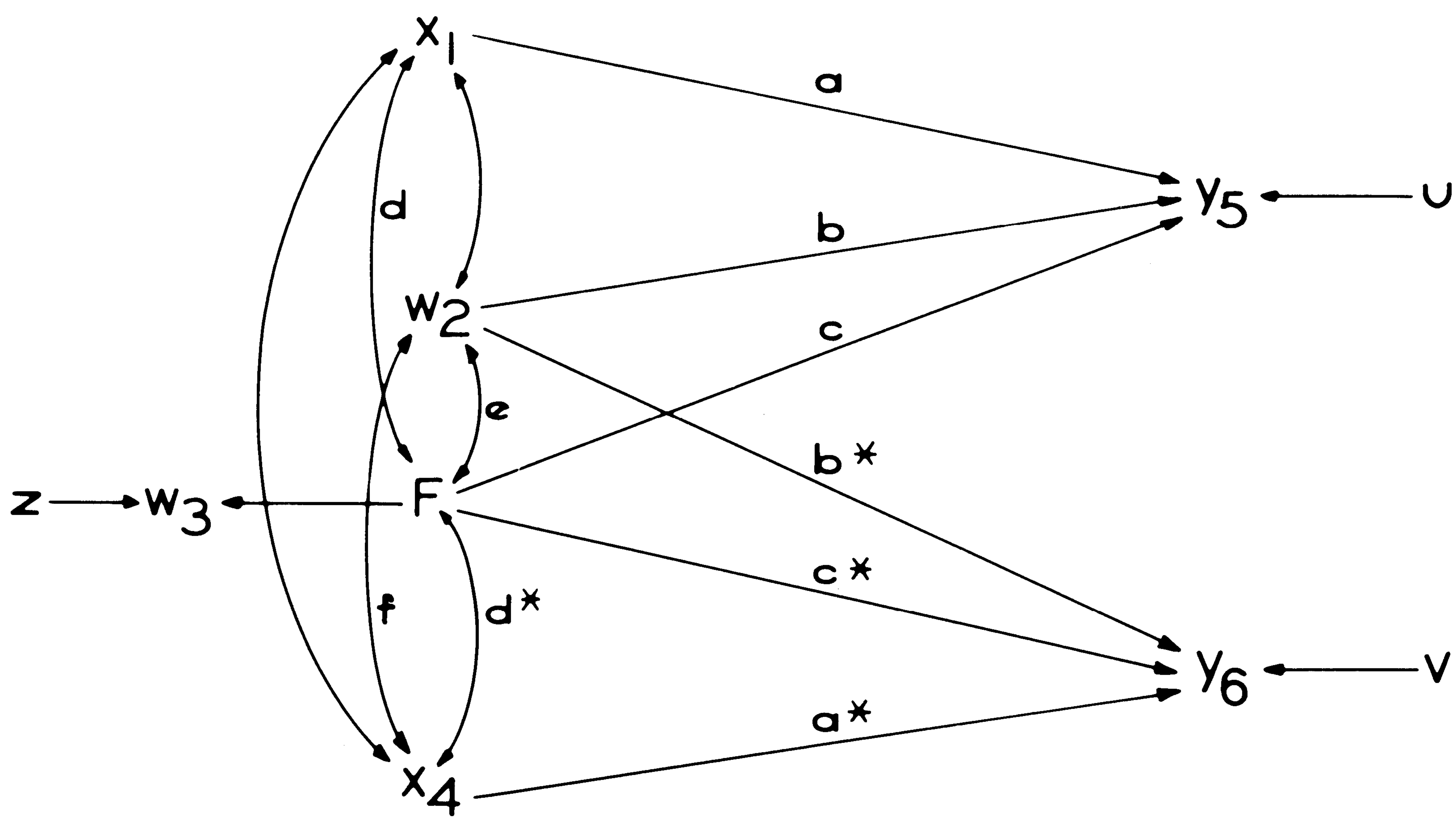
Bureau of the Census. Two results are worth noting here. First, the coefficients of corresponding variables in the two equations are generally similar; probably none of the differences are statistically significant. Again, this is important (for reasons given below) because the results are not based on random pairs of brothers, but rather on men and their oldest brothers. That is, there is some empirical support for the assumption of equality in the effects of family background variables across members of the same family. Of course, we shall be able to test this assumption directly in the Wisconsin data.

Second, in Figure 2 there is a large correlation between the disturbances in the schooling equations. This correlation, $r_{uv} = .437$, is about two-thirds as large as the observed correlation between the schooling of men and their brothers, $r_{56} = .576$. To put the matter in a slightly different way, the four measured background variables explain just over half of the observed correlation. There is no reason to believe that the addition of more socioeconomic background measures to the model would substantially improve the explanation of brothers' resemblance in schooling. To give a rough explanation of this, assume the equations in the top and bottom half of Figure 2 have identical coefficients and there are equal variances in the brothers' schooling. In this case the R^2 (proportion of variance explained) in each equation would be identical to the correlation predicted by the model between the schooling of brothers, not including the effect of the correlation between disturbances. However, there is no evidence that shared family background measures, of the sort usually measured by students of social stratification, can explain as much as 58 percent of the variance in schooling. Yet this fraction (58 percent) is the share of variance one would have to explain to account for the correlation ($r_{56} = .576$) between the schooling of brothers. Of course as Sewell and his associates have shown, it is possible to explain 55 to 65 percent of the variance in post-high school education, but only with the use of additional explanatory variables (like ability, grades, and aspirations) on which siblings would not have identical values.

Figure 3 shows a model of sibling resemblance which includes an unmeasured family background factor. It is highly oversimplified and intended only to have illustrative value. As above, y_5 and y_6 are measures of the achievement of respondent and sibling, respectively. There is also a pair of measurements, x_1 and x_4, respectively, of a predetermined variable on which each sibling has a unique value. There are two common measured background variables, w_2 and w_3, on which the siblings have the same value, i.e., variables like the x's in Figure 1 and Figure 2. In addition there is a common causal variable, F, which has not been observed for any respondent or sibling, but whose properties are to be deduced from the postulated causal structure and the values of correlations among observed variables.

The causal structure of the model in Figure 3 differs in three important ways from that of the two preceding models. First, we show no

Figure 3: A path model of the effects of measured and unmeasured family background variables and sibling characteristics on the achievement of siblings



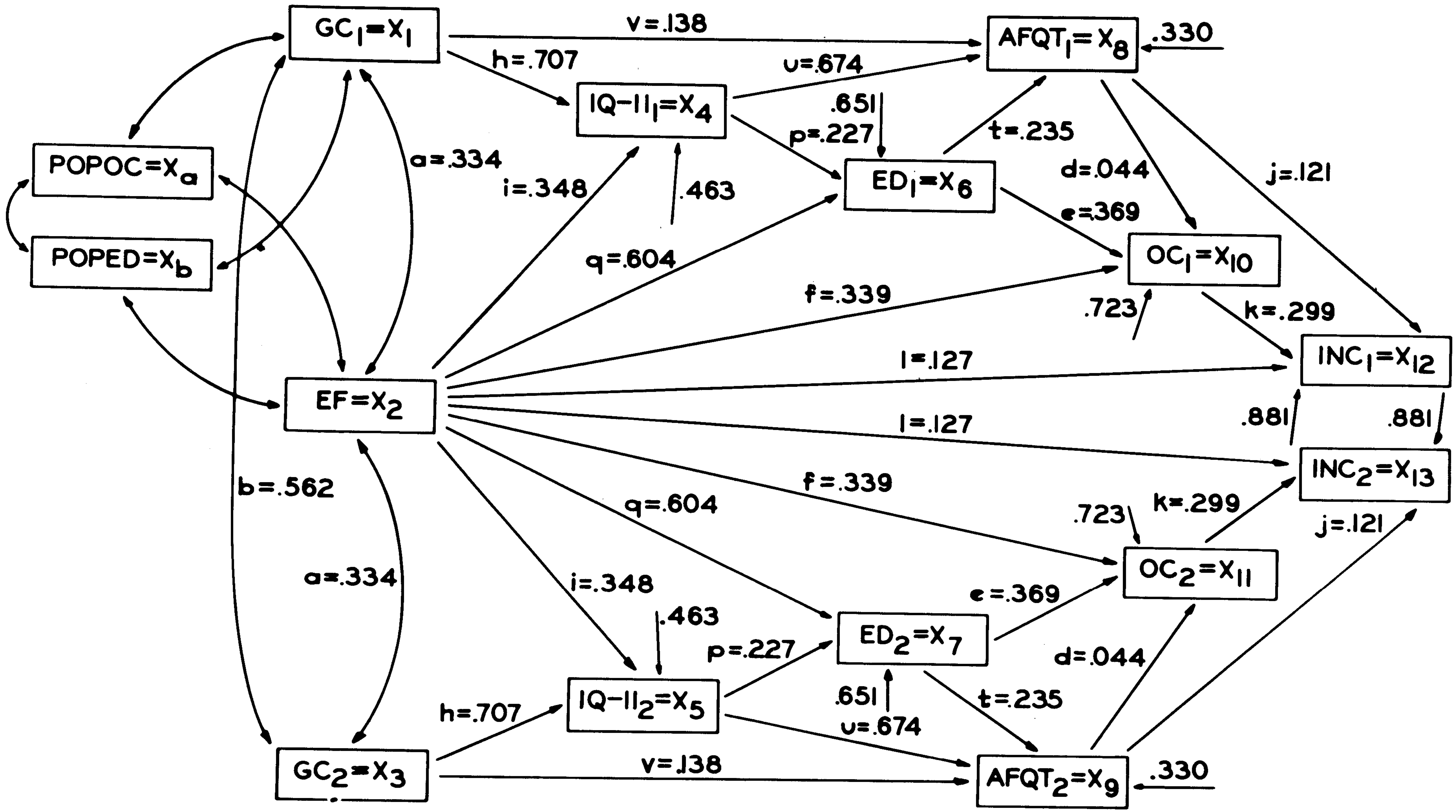
NOTE: x_1 and y_5 are characteristics of a respondent; x_4 and y_6, respectively, are the same characteristics of a sibling; w_2 and w_3 are common measured background variables; F is an hypothetical construct representing unmeasured family influences.

correlation between the disturbances, u and v, of the two endogenous achievement variables, y_5 and y_6. That is, we assume that the specification of a single common family factor, F, is sufficient (in combination with measured family background variables and other sibling characteristics) to account fully for the resemblance of siblings in respect to y. Second, while two of the common background variables (w_2 and F) each affect both y_5 and y_6, the two measures of x affect only the achievement of the sibling to whom they pertain. That is, x_1 affects y_5, but not y_6, and x_4 affects y_6, but not y_5. Third, while one of the common measured family background variables (w_2) affects the y's directly, the other (w_3) does not. Rather, w_3 is correlated with other variables in the model only by virtue of its association with the unmeasured family factor, F.

What substantive notions underly this specification? Suppose that y_5 and y_6 are measures of educational attainment, x_1 and x_4 are measures of academic ability, w_3 is a composite measure of the socioeconomic standing of the family of orientation and w_2 is a measure of the size of the family of orientation. Then the model says that the educational attainment of each sibling is a function of the size of his family of orientation and of another family factor associated with it, as well as of his realized ability. However, the social standing of the family of orientation does not affect schooling directly, but is associated with it only because it reflects more proximate features of family organization and interaction, represented by the unmeasured variable, F. Moreover, there is no reason to expect that the ability of one sibling will affect the schooling of the other. Note there are two very strong assumptions underlying this illustration: first, that some measured variables affect the achievement of one sibling, but not the other, and second, that some variables common to the siblings do not directly affect their achievement outcomes. It is relatively easy to think of variables which have the first of these properties. For example, there is no reason to expect the schooling of one sibling to affect the occupation or earnings of another. It is more difficult to decide which common background variables might be expected to affect achievements directly, and which might not. However, such assumptions are fundamental in enabling us to identify the more plausible causal paths in the system.

Each of the unknown path coefficients or correlations in the model of Figure 3 has been denoted by a lower case letter, and starred and unstarred symbols have been used to denote the corresponding coefficients for the members of a sibling pair. (We ignore the three path coefficients of disturbances, p_{5u}, p_{6v}, and p_{3z}, which can be identified from the equations for complete determination of y_5, y_6, and w_3, once the other coefficients are known.) There are two interesting cases in which to consider the identification of the model of Figure 3. In the first case all of the 15 correlations among the six observed variables are taken to be distinct; that is, we impose no conditions of symmetry or equality between the non-zero coefficients pertaining to the two members of a sibling pair. In this case three correlations among the exogenous variables are fixed as observed (r_{12}, r_{14}, r_{23}), and we estimate six effects of the exogenous variables (a, b, c, a*, b*, c*) and four other correlations involving F (d, d*, e, f). There are a total of 13 correlations or effects to be estimated from 15 correlations among observed variables. The excess of observed correlations is not a sufficient condition for there to exist a solution for each unknown parameter, but in the present case such a solution does exist. (In fact there is more than one solution; the model is overidentified.) In the other interesting case we assume a population of sib pairs drawn strictly at random from a sample of families, so there must be perfect symmetry (equality) in the results for each member of a pair. That is, a = a*, b = b*, c = c*, and d = d*. While we have fewer parameters to estimate in this case, we also have fewer distinct observed correlations. That is, $r_{15} = r_{46}$, $r_{16} = r_{45}$, $r_{12} = r_{42}$, $r_{13} = r_{43}$, $r_{52} = r_{62}$, and $r_{53} = r_{63}$, so there are only nine distinct elements in the correlations among the six measured variables. However, there are not only two distinct unanalyzed correlations among exogenous variables ($r_{12} = r_{24}$, r_{14}), three distinct effects of the exogenous variables (a = a*, b = b*, c = c*), and three distinct correlations involving F (d = d*, e, f). Again, in this case there is more than enough information to permit estimation of each of the parameters of the model. We shall not pursue this example further, but simply close by reiterating that the model of Figure 3 raises issues of specification and identification which will recur throughout our analyses of sibling resemblance.

Just how elaborate are these models of sibling resemblance likely to become? Figure 4 shows a path model of the achievements of U.S. men and their brothers which Hauser and Dickinson (1974) constructed from a set of correlations assembled by Jencks, et al. (1972) from diverse data sources. Neither the specification of the model nor the numbers should be taken seriously except as an indication of the degree of elaboration toward which we shall be aiming in our analyses of sibling resemblance. Because of limitations in the available data, the model of Figure 4 assumes the identity of corresponding coefficients of brothers throughout the model. We shall not have to be quite so restrictive to obtain estimates in the Wisconsin data. Also, the model of Figure 4 does not permit the measured common family background characteristics (x_a and x_b) to affect any measured outcome variables; x_1, x_2, and x_3 are each unobserved variables. As indicated above, we shall permit both measured and unmeasured family characteristics to affect outcomes directly. At the same time we plan to be less ambitious in offering interpretations of the relative importance of heredity and environment. That issue is


GC1=X1
POPOC=Xa
POPED=Xb
EF=X2
GC2=X3
IQ-II1=X4
IQ-II2=X5
ED1=X6
ED2=X7
AFQT1=X8
AFQT2=X9
OC1=X10
OC2=X11
INC1=X12
INC2=X13
v=.138
h=.707
u=.674
.330
.651
p=.227
t=.235
a=.334
i=.348
.463
d=.044
j=.121
e=.369
q=.604
f=.339
.723
k=.299
l=.127
.881
b=.562

treated explicitly in the model of Figure 4, but we do not believe there is any relevant information about it in the Wisconsin data.

One substantive observation about the model in Figure 4 seems appropriate. Even after such intervening outcomes as realized ability and schooling have been taken into account, the family of orientation has persistent effects on occupation and income. Such fundamental tendencies toward the stratification of opportunities across generations are likely to escape detection in a sample of persons, but not in a sample of families. In a cross-section of persons one could not estimate the effect of family of orientation, nor even discover the limits of a proposed explanation which did not include that effect. For example, the typical "status attainment" model looks like the upper or the lower half of Figure 1 or Figure 2, and half the model would not tell you that the model could not account for the resemblance of siblings. This is the reason efforts to measure and interpret sibling resemblance are of such great importance in the study of social stratification and inequality.

REFERENCES

Adams, Bert N. "Birth order: a critical review." Sociometry 35 (September): pp. 411-439, 1972.

Adams, Bert N. and Miles T. Meidam. "Economics, family structure, and college attendance." American Journal of Sociology 74 (November): pp. 230-239, 1968.

Altus, William D. "Birth order and its sequelae." Science 151 (January): pp. 44-48, 1966.

Alwin, Duane F. and Robert M. Hauser. "The decomposition of effects in path analysis." American Sociological Review 40 (February): pp. 37-47, 1975.

Bayer, Alan E. and John K. Folger. "The current state of birth order research." International Journal of Psychiatry 3 (January): pp. 37-39, 1967.

Blau, Peter M. and Otis Dudley Duncan. The American Occupational Structure. New York: John Wiley, 1967.

Bowles, Samuel. "Schooling and inequality from generation to generation." Journal of Political Economy 80 (May/June, Part II): S219-S251, 1972.

Bradley, Richard W. "Birth order and school-related behavior: a heuristic review." Psychological Bulletin 70: pp. 45-51, 1968.

Clarridge,Brian R., Linda S. Sheehy and Taissa S. Hauser. "Tracing members of a panel: a 17-year follow-up." Center for Demography and Ecology Working Paper 76-23, University of Wisconsin-Madison, 1976.

Duncan, Otis D., David L. Featherman, and Beverly Duncan. Socioeconomic Background and Achievement. New York: Seminar Press, 1972.

Featherman, David L. and Robert M. Hauser. "Design for a replicate study of social mobility in the United States." Pp. 219-252 in Kenneth C. Land and Seymour Spilerman (eds.), Social Indicator Models, New York: Russell Sage, 1972.

Hauser, Robert M. Socioeconomic Background and Educational Performance. Rose Monograph Series, American Sociological Association, Washington, D.C., 1971.

Hauser, Robert M., and Thomas N. Daymont. "Schooling, ability and earnings: cross-sectional findings 8 to 14 years after high school graduation." CDE Working Paper 76-19, University of Wisconsin-Madison, 1976.

Hauser, Robert M., and Peter J. Dickinson. "Inequality on occupational status and income." American Educational Research Journal 11 (Spring): pp. 161-168, 1974.

Hauser, Robert M., and David L. Featherman. "Equality of schooling: trends and prospects." Sociology of Education 49 (April): pp. 99-120, 1976.

Hauser, Robert M., William H. Sewell, and Duane F. Alwin. "High school effects on achievement." Pp. 309-341 in William H. Sewell, Robert M. Hauser, and David L. Featherman (eds.), Schooling and Achievement in American Society, New York: Academic Press, 1976.

Hermalin, Albert I. The Homogeneity of Siblings on Education and Occupation. Unpublished Ph.D. dissertation, Princeton University, 1969.

Jencks, Christopher S., et al. Inequality: A Reassessment of the Effect of Family and Schooling in America. New York: Basic Books, 1972.

Lindert, Peter. "Family inputs and inequality among children." Institute for Research on Poverty Discussion Paper 218-74. University of Wisconsin-Madison, 1974.

Mason, William M., Robert M. Hauser, Alan C. Kerckhoff, Sharon S. Poss, and Kenneth Manton. "Models of response error in student reports of parental socioeconomic characteristics." Pp. 443-494 in William H. Sewell, Robert M. Hauser and David L. Featherman (eds.), Schooling and Achievement in American Society, New York: Academic Press, 1976.

Olneck, Michael R. The Determinants of Educational Attainment and Adult Status Among Brothers: The Kalamazoo Study. Unpublished doctoral dissertation, Harvard University, 1976.

Schooler, Carmi. "Birth order effects: not here, not now!" Psychological Bulletin 78: pp. 161-175, 1972.

Sewell, William H. "Community of residence and college plans." American Sociological Review 29 (February): pp. 24-38, 1964.

Sewell, William H. and J. Michael Armer. "Neighborhood context and college plans." American Sociological Review 31 (April): pp. 159-168, 1966.

Sewell, William H., Archibald O. Haller, and Alejandro Portes. "The educational and early occupational attainment process." American Sociological Review 34 (February): pp. 82-92, 1969.

Sewell, William H., Archibald O. Haller, and George W. Ohlendorf. "The educational and early occupational status attainment process: replication and revision." American Sociological Review 35 (December): pp. 1014-1027, 1970.

Sewell, William H. and Robert M. Hauser. Education, Occupation, and Earnings: Achievement in the Early Career. New York: Academic Press, 1975.

Sewell, William H. and Alan M. Orenstein. "Community of residence and occupational choice." American Journal of Sociology, 70 (March): pp. 551-563, 1965.

Sewell, William H. and Vimal P. Shah. "Socioeconomic status, intelligence, and the attainment of higher education." Sociology of Education, 40 (Winter): pp. 1-23, 1967.

Sewell, William H. and Vimal P. Shah. "Social class, parental encouragement, and educational aspirations." The American Journal of Sociology, 73 (March): pp. 559-572, 1968.

Zajonc, R.M. and G.B. Markus. "Birth order and intellectual development." Psychological Review 82: pp. 74-88, 1975.

COMMENTS ON W. SEWELL'S AND R. HAUSER'S "ON THE EFFECTS OF FAMILIES AND FAMILY STRUCTURE"

Ann P. Bartel
Assistant Professor of Economics, University of Pennsylvania

In this paper, William Sewell and Robert Hauser discuss several methods by which the Wisconsin Data can be used to measure the effect of family background on an individual's educational attainment and post-school achievement. This data set has several features which make it quite attractive for studying the effects of families. First, it contains information on parental income at the time the respondent was a high-school senior. This figure was obtained directly from tax records and is, therefore, relatively free of reporting error. Second, there is information on both male and female siblings. It is therefore possible to separately estimate family effects for same-sex and cross-sex pairs. Finally, the collection of data on the Wisconsin siblings is still in progress and we have the rare opportunity of ensuring that all the necessary variables are obtained.

In my discussion, I would like to suggest some additional ways in which the Wisconsin Data can be used to study the effects of families as well as to point out the additional sibling variables that should be obtained. Following Sewell and Hauser's format, I will first discuss the problem of explaining sibling differences and then handle the question of why siblings resemble one another.

Sibling Differences

As Sewell and Hauser suggest, within-family regressions can be used to estimate the effect of birth order, net of the family effect. This procedure would enable us to determine if, for example, the first-born in the family acquires more schooling and if, holding schooling constant, he does better in the labor market than a younger sibling. Sewell and Hauser do not, however, mention the importance of holding certain variables constant in order to obtain the true birth order effects in these regressions. For example, in the schooling equation, one should standardize for the differences in ability between the respondent and his sibling. The first-born child could have a higher pre-school ability because his parents spent more time with him and this higher ability could explain his higher educational attainment. In order to examine whether birth order has an effect on earnings, one needs to hold constant differences in education, differences in labor force experience, and differences in hours worked per week. It will be interesting to see if birth order has any effect beyond educational attainment. Thus, in order to correctly perform this analysis, data on the earnings, hours and labor force experience of the randomly chosen siblings must be collected. Similar procedures can be used to estimate the effects of sex, birth year and spacing as well as to test for the presence of interaction effects between, for example, sex and birth order. In terms of analyzing the effect of birth year, it may only be necessary to distinguish between individuals born prior to World War II and those born during and after the war years.

The within-family regressions also provide an opportunity to estimate the rate of return to schooling, net of the common genetic structure and common family background that the siblings share. This coefficient should then be compared to the schooling coefficients obtained from two earnings functions estimated across individuals where one holds family income and other measured family background variables constant and the other does not. This should enable us to see to what extent the measured family variables serve as a proxy for the true family effect. These comparisons should be done separately for male pairs, female pairs and cross-sex pairs. It will also be interesting to see how these comparisons relate to similar tests done by Behrman, Taubman and Wales on dizygotic twins and by Chamberlain and Griliches on brothers.

Sibling Resemblance

In order to determine why siblings are more alike than unrelated persons, Sewell and Hauser propose an analysis of sibling correlations. My basic suggestion for this analysis is that, prior to employing any latent variable techniques, Sewell and Hauser should focus on utilizing the measured family background variables, in particular family income, to explain the observed sibling correlations. The Wisconsin Data enable one to estimate the contributions of family income, father's education and occupation, mother's education and labor force experience, and sex of family head on the sibling correlations of education, occupation, fertility behavior, hours of work and earnings. It will be interesting to see how the contributions differ for each of the observed sibling correlations as well as how the estimated contributions are affected if one looks at same-sex pairs or cross-sex pairs. Of course, in explaining the sibling correlation in earnings, it will be important to control for possible similarities in education, experience and hours of work.

The Wisconsin data set should also prove to be an excellent one for the use of latent variable techniques. It would be particularly interesting to test a model similar to the one utilized by Chamberlain and Griliches for the NLS brothers. The Wisconsin data set has a major advantage over the NLS data in that actual, rather than expected, earnings would be available for the respondent and his sibling. Moreover, it would be possible to see if the contribution of unmeasured family background differs if one has information on parental income. The wealth of information available in the Wisconsin Data should enable Sewell and Hauser to estimate a latent variable model that does not require many of the restrictive assumptions that have been necessary with other data sets.

CHAPTER 9

ARE BROTHERS AS GOOD AS TWINS?

Gary Chamberlain*
Harvard University

1. Introduction

We shall try to determine whether twin data can identify more general models than can be identified just using brothers. The issue is not whether a twin sample will give more precise estimates (for a given sample size), but whether it will allow us to answer different questions. We want to know if the twin data is more useful in controlling for omitted variables—for example, in controlling for unobserved "ability" in order to obtain unbiased estimates of the effects of schooling on earnings. If the only omitted variables are either genetic or common family background, then within-pair regressions using the monozygotic twins will identify models that cannot be identified using sibs. But if the omitted variables have non-genetic components that vary within families, then a model is identified using data on twins only if it is identified using data just on brothers.

2. The Main Results

The following one-factor model provides a simple illustration of our results:

$$y_{1ij} = \lambda_1 h_{ij} + u_{1ij}$$

$$y_{2ij} = \gamma_1 y_{1ij} + \lambda_2 h_{ij} + u_{2ij}$$

$$y_{3ij} = \gamma_2 y_{1ij} + \gamma_3 y_{2ij} + \lambda_3 h_{ij} + u_{3ij}.$$ [1]

The subscripts refer to the jth individual in the ith family. The common omitted variable h is the only source of cross-equation correlation in the residuals, since the u's are assumed to be uncorrelated with each other.

*I am grateful for discussions with A. Goldberger, Z. Griliches, and P. Taubman. Financial support was provided by the NIE and NSF.

Since h is correlated with y_1 and y_2, estimates of the γ's based on regressions that do not include h are biased. The u's are assumed to be uncorrelated across members within a family: $E(u_{ki\alpha} \; u_{ki\beta}) = 0$ if $\alpha \neq \beta$; hence the cross-sib correlations are generated solely through h. We shall model these correlations by assigning h a variance-components structure:

$$h_{ij} = f_i + g_{ij},$$

where the f_i are independent and identically distributed across families and the g_{ij} are independent and identically distributed within families.[2] Hence the correlation on h between a pair of sibs is

$$\rho_h = E(h_{i\alpha} \; h_{i\beta})/E(h_{i\alpha}^2) = \sigma_f^2/(\sigma_f^2 + \sigma_g^2).$$

An example of the model has y_1 = years of schooling, y_2 = a late (post school) test score, which depends on schooling, and y_3 = earnings. The unobservable h can be interpreted as early "ability". Then we can interpret u_2 as measurement error in the test, in which case a plausible restriction is $\gamma_3 = 0$. This excludes the measured score from the earnings equation since the true score components (y_1 and h) are included.[3]

This model can be applied to sibs, dizygotic (DZ) twins, or monozygotic (MZ) twins. We shall link these three populations by assuming that the structural γ's, λ's, and σ's are the same in all of them. We assume, however, that the sibs and DZ twins differ from the MZ twins in the decomposition of h into betweeen- and within-family components.[4]

We can model these differences by decomposing

$$h_{ij} = G_{ij} + N_{ij},$$

where G is the genetic component and N is the non-genetic or environmental component. So the family effects f_i include both common genes and common environment. A convenient simplification is to assume a zero correlation between G and N. This can be achieved by interpreting G as the part of h that is (linearly) predictable from genes, with the residual N uncorrelated with G by construction. If instead we carry along a correlation between G and N, then the twin model will be even less identified, strengthening our results.

With G and N uncorrelated, the cross-sib correlation on h decomposes as follows:

$$\rho_h = \psi \rho_G + (1 - \psi) \rho_N,$$

where ρ_G and ρ_N are the cross-sib correlations on G and N, and $\psi = \sigma_G^2/(\sigma_G^2 + \sigma_N^2)$. For the MZ twins, $\rho_G = 1$. If there is no dominance, epistasis, or assortative mating, then $\rho_G = .5$ for the DZ twins.[5] Also we might constrain ρ_N and ψ to be the same for both twin types. But ρ_h will still remain unconstrained for both MZ and DZ twins, since in general ρ_N and ψ are unknown.

The reduced form of the model is

$$y_k = d_k h + e_k, \qquad k = 1, 2, 3,$$

where

$$\underset{\sim}{d} = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix} = \begin{bmatrix} \lambda_1 \\ \lambda_2 + \gamma_1 \lambda_1 \\ \lambda_3 + \gamma_2 \lambda_1 + \gamma_3(\lambda_2 + \gamma_1 \lambda_1) \end{bmatrix},$$

and

$$\underset{\sim}{e} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 + \gamma_1 u_1 \\ u_3 + \gamma_2 u_1 + \gamma_3(u_2 + \gamma_1 u_1) \end{bmatrix}.$$

Under normality, or limiting ourselves to second order moments, the distribution of $\underset{\sim}{y}' = (y_1, y_2, y_3)$ is completely characterized by the following covariance matrices (we are suppressing the intercepts and assuming that all means are zero):

$$\underset{\sim}{\Omega} = E(\underset{\sim}{y}_{ij} \underset{\sim}{y}'_{ij}) = \underset{\sim}{d} \underset{\sim}{d}' + \underset{\sim}{V}$$

$$\underset{\sim}{\Theta} = E(\underset{\sim}{y}_{i\alpha}\, \underset{\sim}{y}'_{i\beta}) = \underset{\sim}{d}\, \underset{\sim}{d}'\rho_h, \qquad \beta \neq \alpha ,$$

$$\underset{\sim}{\Sigma} = \underset{\sim}{\Omega} - \underset{\sim}{\Theta} = \underset{\sim}{d}\, \underset{\sim}{d}' (1 - \rho_h) + \underset{\sim}{V},$$

where

$$\underset{\sim}{V} = E(\underset{\sim}{e}_{ij}\, \underset{\sim}{e}'_{ij}) =$$

$$\begin{bmatrix} \sigma_1^2 & \gamma_1 \sigma_1^2 & (\gamma_2 + \gamma_3 \gamma_1)\sigma_1^2 \\ & \sigma_2^2 + \gamma_1^2 \sigma_1^2 & \gamma_3 \sigma_2^2 + \gamma_1(\gamma_2 + \gamma_3 \gamma_1)\sigma_1^2 \\ & & \sigma_3^2 + (\gamma_2 + \gamma_3 \gamma_1)^2 \sigma_1^2 + \gamma_3^2 \sigma_2^2 \end{bmatrix},$$

and we have set the scale of the unobservable by the normalization $\sigma_h = 1$. $\underset{\sim}{\Theta}$ is the cross-sib covariance matrix, generated by the family effects f_i; $\underset{\sim}{\Sigma}$ is the within-family covariance matrix, generated by the individual effects g_{ij} and by the disturbances u_{ij}; and $\underset{\sim}{\Omega} = \underset{\sim}{\Theta} + \underset{\sim}{\Sigma}$ is the total covariance matrix, including both family and individual effects.

The identification problem in the sib model is to solve for the structural parameters from $\underset{\sim}{\Omega}$ and $\underset{\sim}{\Theta}$ (or $\underset{\sim}{\Sigma}$). A simple count of the unknown parameters gives 3 $\tilde{\gamma}$'s, 3 $\tilde{\lambda}$'s, 3 $\tilde{\sigma}$'s, and ρ_h for a total of 10 parameters. The symmetric matrices $\underset{\sim}{\Omega}$ and $\underset{\sim}{\Theta}$ contain 12 distinct elements, and so they appear to generate enough equations. $\underset{\sim}{\Theta}$, however, is constrained to have rank 1, implying that it has only 3 free elements. So there are 10 unknown parameters and only 9 equations; hence the sib model is not identified without an additional restriction, such as $\gamma_3 = 0$.[6]

In the twin model there is a cross-sib covariance matrix for the DZ pairs:

$$\underset{\sim}{\Theta}_D = \underset{\sim}{d}\, \underset{\sim}{d}'\rho_{hD},$$

a cross-sib covariance matrix for the MZ pairs:

$$\underset{\sim}{\Theta}_M = \underset{\sim}{d}\, \underset{\sim}{d}'\rho_{hM},$$

and a total covariance matrix:

$$\underset{\sim}{\Omega} = \underset{\sim}{d}\,\underset{\sim}{d}' \qquad + \underset{\sim}{V},$$

which is the same for both twin types. The two twin types differ only in the decomposition of $\underset{\sim}{\Omega}$ into between- and within-family components.

The additional set of cross-sib covariances might appear to provide enough degrees of freedom. Even given the rank = 1 constraints on the $\underset{\sim}{\Theta}$'s, three additional equations would be more than enough, since there is only one additional parameter,due to a separate ρ_{hD} and ρ_{hM} for the DZ and MZ twins. But there are constraints across $\underset{\sim}{\Theta}_D$ and $\underset{\sim}{\Theta}_M$: $\underset{\sim}{\Theta}_M = \frac{\rho_{hM}}{\rho_{hD}} \underset{\sim}{\Theta}_D$; hence the additional set of cross-sib covariances generates only a single additional equation, giving 10 equations and 11 unknowns. Thus the twin model is just as underidentified as the sib model; in this example they each need one additional restriction.

An exception to this result occurs if $\rho_h = 1$ for the MZ twins. Then there are only 10 unknowns, and the model can be identified by using the MZ twins to form the differences within the pairs on each of the variables: $y_{ki\alpha} - y_{ki\beta}$. Regressions based on these within-pair differences will give unbiased estimates since genetic variables are eliminated. But in general we should not impose a priori that there is no within-family variation in the omitted environmental variables. If we do not make this assumption, then the conditions for identification in the twin model are identical to the conditions in the sib model.

3. <u>Extensions</u>

The extension to more equations is straightforward. In the sib model with m equations, there are m(m-1)/2 γ's (still assuming, for the moment, a triangular structure), m λ's, m σ's, and ρ_h for a total of m(m + 1)/2 + m + 1 structural parameters. Counting up the unrestricted elements in the reduced form gives m(m + 1)/2 from $\underset{\sim}{\Omega}$ and m from the rank one $\underset{\sim}{\Theta} = \underset{\sim}{d}\,\underset{\sim}{d}'\rho_h$, for a total of m(m + 1)/2 + m; hence one additional restriction is still needed for identification.

In the twin model there are m(m + 1)/2 + m + 2 parameters, due to the separate ρ_h's for DZ and MZ twins. In the reduced form, $\underset{\sim}{\Theta}_D = \underset{\sim}{d}\,\underset{\sim}{d}'\rho_{hD}$ is proportional to $\underset{\sim}{\Theta}_M = \underset{\sim}{d}\,\underset{\sim}{d}'\rho_{hM}$; hence the additional set of cross-sib covariances adds only one more equation. So there are m(m + 1)/2 + m + 1 equations, and once again the degree of underidentification is precisely the same as in the sib model.

In the non-triangular case, it is still true that the reduced form of the twin model has only one more free parameter than the reduced form of the sib model. For the reduced form can still be written as $\underset{\sim}{y} = \underset{\sim}{d}\;h + \underset{\sim}{e}$, and so $\underset{\sim}{\Theta}_M = \underset{\sim}{d}\,\underset{\sim}{d}'\rho_{hM}$ is still proportional to

$\underset{\sim}{\theta}_D = \underset{\sim}{d}\,\underset{\sim}{d}'\,\rho_{hD}$. Since the twin model still has one more structural parameter, the degree of over- or underidentification is the same as in the sib model.

Finally, we want to extend our results to the general n-factor case. The notation is

$$\underset{\sim}{\Gamma}\,\underset{\sim}{y}_{ij} = \underset{\sim}{\Lambda}\,\underset{\sim}{h}_{ij} + \underset{\sim}{u}_{ij}$$

$$\underset{\sim}{h}_{ij} = \underset{\sim}{f}_{i} + \underset{\sim}{g}_{ij},$$

where $\underset{\sim}{\Gamma}$ is an m x m matrix of structural coefficients (the γ's), $\underset{\sim}{y}$ is an m x 1 vector of endogenous variables, $\underset{\sim}{\Lambda}$ is an m x n matrix of factor coefficients, $\underset{\sim}{h}$ is an n x 1 vector of unobserved factors, and $\underset{\sim}{u}$ is an m x 1 vector of structural disturbances; $\underset{\sim}{f}$ is an n x 1 vector of family effects, including common family background and common genes, and $\underset{\sim}{g}$ is an n x 1 vector of individual effects. Let

$$E(\underset{\sim}{f}\,\underset{\sim}{f}') = \underset{\sim}{\phi}_f, \quad E(\underset{\sim}{g}\,\underset{\sim}{g}') = \underset{\sim}{\phi}_g,$$

and

$$\underset{\sim}{\phi}_f + \underset{\sim}{\phi}_g = \underset{\sim}{\phi}_h.$$

We assume that $\underset{\sim}{\phi}_h$ is the same for both twin types, whereas the decomposition into $\underset{\sim}{\phi}_f$ and $\underset{\sim}{\phi}_g$ is different for the DZ and the MZ twins.

The reduced form is

$$\underset{\sim}{y} = \underset{\sim}{D}\,\underset{\sim}{h} + \underset{\sim}{e},$$

where $\underset{\sim}{D} = \underset{\sim}{\Gamma}^{-1}\underset{\sim}{\Lambda}$ is a matrix of reduced form coefficients, and $\underset{\sim}{e} = \underset{\sim}{\Gamma}^{-1}\underset{\sim}{u}$ is a vector of reduced form disturbances. The reduced form covariances are

$$\underset{\sim}{\Omega} = \underset{\sim}{D}\,\underset{\sim}{\phi}_h\,\underset{\sim}{D}' + \underset{\sim}{V}$$

$$\underset{\sim}{\Theta}_D = \underset{\sim}{D}\,\underset{\sim}{\phi}_{fD}\,\underset{\sim}{D}'$$

$$\underset{\sim}{\Theta}_M = \underset{\sim}{D}\,\underset{\sim}{\phi}_{fM}\,\underset{\sim}{D}',$$

where $\underset{\sim}{V} = E(\underset{\sim}{e}\,\underset{\sim}{e}') = \underset{\sim}{\Gamma}^{-1}\underset{\sim}{U}\,\underset{\sim}{\Gamma}^{-1}$ and $\underset{\sim}{U} = E(\underset{\sim}{u}\,\underset{\sim}{u}') = \text{diag}\{\sigma_1^2, \ldots, \sigma_M^2\}$. The additional information in the twin model comes from the additional set of cross-sib covariances. If, however, there are no restrictions on $\underset{\sim}{\phi}_h$, $\underset{\sim}{\phi}_{fD}$, or $\underset{\sim}{\phi}_{fM}$, then we can show that the additional set of cross-sib covariances does not help to identify the γ's.

In order to consider the potential restrictions on $\underset{\sim}{\phi}_h$, $\underset{\sim}{\phi}_{fD}$, and

$\underset{\sim}{\phi}_{fM}$, we decompose $\underset{\sim}{h}$ into $\underset{\sim}{h} = \underset{\sim}{G} + \underset{\sim}{N}$, where $\underset{\sim}{G}$ is a vector of genetic effects and $\underset{\sim}{N}$ is a vector of environmental effects. As before we assume that $\underset{\sim}{G}$ and $\underset{\sim}{N}$ are uncorrelated; allowing for correlation would only strengthen our results.

Then the factor covariances are

$$\underset{\sim}{\phi}_h = E(\underset{\sim}{G}_\alpha \underset{\sim}{G}'_\alpha) + E(\underset{\sim}{N}_\alpha \underset{\sim}{N}'_\alpha)$$

$$\underset{\sim}{\phi}_f = E(\underset{\sim}{G}_\alpha \underset{\sim}{G}'_\beta) + E(\underset{\sim}{N}_\alpha \underset{\sim}{N}'_\beta), \qquad \alpha \neq \beta .$$

For the MZ twins, we have $E(\underset{\sim}{G}_\alpha \underset{\sim}{G}'_\beta) = E(\underset{\sim}{G}_\alpha \underset{\sim}{G}'_\alpha)$. A very restrictive assumption on the genetic components for the DZ twins is that $E(\underset{\sim}{G}_\alpha \underset{\sim}{G}'_\beta) = .5\ E(\underset{\sim}{G}_\alpha \underset{\sim}{G}'_\alpha)$; a less restrictive assumption would only strengthen our results.

These assumptions imply that

$$\underset{\sim}{\phi}_{fD} = .5\ E(\underset{\sim}{G}_\alpha \underset{\sim}{G}'_\alpha) + E(\underset{\sim}{N}_\alpha \underset{\sim}{N}'_\beta)$$

$$\underset{\sim}{\phi}_{fM} = E(\underset{\sim}{G}_\alpha \underset{\sim}{G}'_\alpha) + E(\underset{\sim}{N}_\alpha \underset{\sim}{N}'_\beta).$$

Since $\underset{\sim}{\phi}_h$, $\underset{\sim}{\phi}_{fD}$, and $\underset{\sim}{\phi}_{fM}$ depend on different linear combinations of $E(\underset{\sim}{G}_\alpha \underset{\sim}{G}'_\alpha)$, $E(\underset{\sim}{N}_\alpha \underset{\sim}{N}'_\alpha)$, and $E(\underset{\sim}{N}_\alpha \underset{\sim}{N}'_\beta)$, we conclude that $\underset{\sim}{\phi}_h$ and the $\underset{\sim}{\phi}_f$'s are unconstrained if there are no restrictions on the decomposition of the environmental covariances into between- and within-family components. In general there will not be any restrictions of this kind, and so $\underset{\sim}{\phi}_h$, $\underset{\sim}{\phi}_{fD}$, and $\underset{\sim}{\phi}_{fM}$ will be unconstrained.

In that case, the availability of $\underset{\sim}{\Theta}_M$ in addition to $\underset{\sim}{\Omega}$ and $\underset{\sim}{\Theta}_D$ (which distinguishes the twin model from the sib model) does not help to identify the γ's. At most $\underset{\sim}{\Theta}_M$ allows us to solve for $\underset{\sim}{\phi}_{fM}$, without helping us to extract $\underset{\sim}{\Gamma}$ from $\underset{\sim}{\Omega}$ and $\underset{\sim}{\Theta}_D$.

Note first that $\underset{\sim}{\Theta}_D$ and $\underset{\sim}{\Theta}_M$ have the same column spaces. So given a decomposition of $\underset{\sim}{\Theta}_D$ into $\underset{\sim}{\Theta}_D = \underset{\sim}{S}\ \underset{\sim}{S}'$, where $\underset{\sim}{S}$ is m x n, we can find an n x n matrix $\underset{\sim}{A}$ such that $\underset{\sim}{\Theta}_M = \underset{\sim}{S}\,\underset{\sim}{A}\,\underset{\sim}{A}'\,\underset{\sim}{S}'$. Thus $\underset{\sim}{\Theta}_M$ adds only n(n + 1)/2 degrees of freedom to the reduced form, corresponding to the n(n + 1)/2 free elements in the symmetric matrix $\underset{\sim}{A}\ \underset{\sim}{A}'$.

The n(n + 1)/2 degrees of freedom in $\underset{\sim}{\ }_M$ will at most allow us to solve for the n(n + 1)/2 elements in $\underset{\sim}{\phi}_{fM}$. So if the $\underset{\sim}{\phi}$'s are unrestricted, then $\underset{\sim}{\Theta}_M$ provides no information on $\underset{\sim}{\phi}_{fD}$ or $\underset{\sim}{\phi}_h$. The problem of extracting $\underset{\sim}{\Gamma}$ from $\underset{\sim}{\Omega}$ and $\underset{\sim}{\Theta}_D$ is unaffected by $\underset{\sim}{\Theta}_M$, since the additional set of cross-sib covariances does not provide any of the structural parameters that determine $\underset{\sim}{\Omega}$ and $\underset{\sim}{\Theta}_D$. We conclude that the additional cross-sib covariances do not add to the identifiability of the γ 's.

As in our initial example, an exception to this result occurs if the only non-genetic omitted variables are common family background. In that case $E(\underset{\sim}{N}_\alpha \underset{\sim}{N}'_\alpha) = E(\underset{\sim}{N}_\alpha \underset{\sim}{N}'_\beta)$, and so the $\underset{\sim}{\phi}$'s are restricted. Then the γ 's can be identified through regressions based on the within-pair differences for the MZ twins. But in general we should not impose a priori that there is no within-family variation in the non-genetic omitted variables.

4. Conclusion

Our framework has been a simultaneous equations model. Its novel feature is the structure of the residuals. This structure combines factor analysis with a variance-components specification. Common left-out variables motivate the factor analysis approach; the availability of family groupings allows us to impose a variance-components structure on the left-out variables. The problem is to combine these residual covariance restrictions with the more conventional slope restrictions in order to identify the structural coefficients.

Our question is whether the identification conditions in the twin model are less restrictive than in the sib model. We are able to answer this without specifying fully the identification conditions, although a complete analysis is available for the triangular case.[7]

The apparent advantage of the twin model is that it has an additional set of cross-sib covariances. These reduced form covariances appear to generate more than enough equations to solve for the structural parameters. But there are constraints connecting the cross-sib covariances for the DZ and MZ twins. In the one-factor case the two covariance matrices are proportional to each other; in general they have identical column spaces. These constraints reduce the effective degrees of freedom in the twin model so that the degree of over- or underidentification is precisely the same as in the sib model.

An exception to this result occurs if the only omitted variables are either genetic or common family background. Then the within-pair regressions for the MZ twins will identify models that cannot be identified just using sibs (provided that there is genetic variation in the omitted variables). In general, however, we should not assume a priori that there is no within-family variation in the left-out environmental variables. If we do not make this assumption, then brothers are as good as twins.

FOOTNOTES

[1]We have suppressed any exogenous variables in order to concentrate on the novel feature of these models--namely that part of the identification comes from restrictions on the residual covariances. It is straightforward to extend our results to models that include exogenous variables.

[2]Balestra and Nerlove (1966), Wallace and Hussein (1969), Nerlove (1971), Madalla (1971), and others have dealt with the single-equation version of this variance-components model.

[3]See Chamberlain and Griliches (1975, 1976) for empirical applications of such models using data on brothers.

[4]For a critical discussion of twin models, see Goldberger (1976a, b). There are empirical applications in Behrman and Taubman (1976), Taubman (1976), Behrman, Taubman, and Wales (1976), and Jencks and Brown (1976).

[5]See Jencks (1972), Appendix A.

[6]Necessary and sufficient conditions for identification in variance-components models with a triangular structure are given in Chamberlain (1976a). For an instrumental variable interpretation of these conditions, see Chamberlain (1976b).

[7]See Chamberlain (1976a, b).

REFERENCES

Balestra, P. and M. Nerlove, 1966. "Pooling Cross-Section and Time-Series Data in the Estimation of a Dynamic Model: The Demand for Natural Gas," Econometrica, 34(3), pp. 585-612.

Behrman, J. and P. Taubman, 1976. "Intergenerational Transmission of Income and Wealth," American Economic Review, 66, pp. 436-440.

Behrman, J., P. Taubman, and T. Wales, 1976. "Controlling for and Measuring the Effects of Genetics and Family Environment in Equations for Schooling and Labor Market Success." This volume.

Chamberlain, G. 1976a. "Identification in Variance-Components Models," Harvard Institute of Economic Research Discussion Paper No. 486.

__________ 1976b. "An Instrumental Variable Interpretation of Identification in Variance-Components and MIMIC Models." . This volume.

Chamberlain, G. and Z. Griliches, 1975. "Unobservables with a Variance-Components Structure: Ability, Schooling, and the Econornc Success of Brothers," International Economic Review, 16(2), pp. 422-449.

__________, 1976. "More on Brothers." This volume.

Goldberger, A.S., 1976a. "On Jensen's Method for Twins," Educational Psychologist, 12, pp. 79-82.

__________, 1976b. "Twin Methods: A Skeptical View." This volume.

Jencks, C. et al, 1972. Inequality. New York: Basic Books.

Jencks, C. and M. Brown, 1976. "Genes and Social Stratification: A Methodological Exploration with Illustrative Data." This volume.

Madalla, G.S., 1971. "The Use of Variance-Components Models in Pooling

Time-Series and Cross-Section Data," Econometrica, 39(2), pp. 341-358.

Nerlove, M., 1971. "A Note on Error-Components Models," Econometrica, 39(2), pp. 383-396.

Taubman, P., 1976. "Earnings, Education, Genetics, and Environment," Journal of Human Resources, 11(4), pp. 447-461.

Wallace, T.D. and A. Hussein, 1969. "The Use of Error-Components Models in Combining Cross-Section with Time-Series Data," Econometrica, 37(1), pp. 55-72.

CHAPTER 10

TWIN METHODS: A SKEPTICAL VIEW

Arthur S. Goldberger
Professor of Economics, University of Wisconsin-Madison

1. INTRODUCTION

On observed variables, MZs (identical twins) are more similar than DZs (fraternal twins). On the latent variable, genotype, it is known that MZs are more similar than DZs. On the latent variable, environment, it is not known whether MZs are more similar than DZs. Assume that they are not. Then attribute the excess observed resemblance of MZs to their excess genetic resemblance, thus estimating the role of heredity.

That is the gist of the twin method. Since the equal-environmental-correlation assumption is questionable, the estimates produced by the method should be viewed skeptically.[1]

2. A UNIVARIATE MODEL

Consider first the model for a single observed variable. An individual's observed phenotype Y is determined as the sum of two unobserved components, genotype X and environment U:

$$Y = X + U.$$

The two components are uncorrelated so that total phenotypic variance is

$$\sigma_Y^2 = \sigma_X^2 + \sigma_U^2.$$

He is paired with another individual for whom $Y^* = X^* + U^*$. Their phenotypic covariance is

$$\sigma_{YY^*} = \sigma_{XX^*} + \sigma_{UU^*},$$

on the assumption that $\sigma_{XU^*} = 0$. Standardizing all variables we write

---

An earlier version of this paper was presented at the MSSB Kinometrics Conference, held at Williamstown, MA, on May 24-26, 1976. My research has been facilitated by grants from the Institute for Research on Poverty and the Graduate School Research Committee at the University of Wisconsin, the National Science Foundation, and the Center for Advanced Study in the Behavioral Sciences. I have drawn freely on discussions and correspondence with Gary Chamberlain, John Conlisk, Christopher Jencks, Richard Lewontin, Michael Olneck, Paul Taubman, and Terence Wales. Computational assistance on Section 7 was provided by David Stapleton. I am grateful for all this help but retain responsibility for errors.

$$y = hx + eu$$

(1) $$1 = h^2 + e^2$$

$$c = gh^2 + \rho e^2$$

where $y = Y/\sigma_Y$, $x = X/\sigma_X$, $u = U/\sigma_U$,

$$h^2 = \sigma_X^2/\sigma_Y^2 = \text{heritability}$$

$$e^2 = \sigma_U^2/\sigma_Y^2 = \text{environmentability}$$

$$c = \sigma_{YY*}/\sigma_Y^2 = \text{phenotypic correlation}$$

$$g = \sigma_{XX*}/\sigma_X^2 = \text{genotypic correlation}$$

$$\rho = \sigma_{UU*}/\sigma_U^2 = \text{environmental correlation.}$$

Note that the observed correlation c is a weighted average of the unobserved correlations g and ρ, the weights being h^2 and $e^2 = 1 - h^2$.

Introduce the subscripts 1 for MZ and 2 for DZ. Since $g_1 = 1$, we have

(2) $$c_1 = h^2 + \rho_1 e^2$$

(3) $$c_2 = g_2 h^2 + \rho_2 e^2.$$

Thus

$$c_1 - c_2 = (1-g_2) h^2 + (\rho_1 - \rho_2) e^2,$$

or, more compactly,

(4) $$\Delta c = \Delta g \cdot h^2 + \Delta\rho \cdot e^2.$$

Note that the excess phenotypic correlation of MZs is a weighted average of their excess phenotypic correlation and their excess environmental correlation, the weights again being h^2 and $e^2 = 1 - h^2$. Solving (4) for h^2, we get

(5) $$h^2 = (\Delta c - \Delta\rho) / (\Delta g - \Delta\rho).$$

On the right-hand side, there are two unknowns ($\Delta\rho$ & Δg) along with the one datum (Δc), so that h^2 is not identified.

Now assume that the environmental correlation for MZs is the same as for DZs, that is $\Delta\rho = 0$. Then

(6) $$h^2 = \Delta c/\Delta g.$$

Equation (6) was obtained by Jensen (1967). Assigning alternative values to Δg, he calculated alternative estimates of h^2 for a number of traits for which MZ and DZ correlations had been reported. Martin (1975, p. 225) provides a more recent illustration. For a sample of approximately 25 MZs and 35 DZs, 15 years old, in South Australia, he takes $\Delta g = 1/2$, and obtains estimates of heritability for various school subjects, which we report below. [Actually Martin works with variances and covariances, but the essence of his calculation is captured by (6)]. Equation (6) is also Jencks & Brown's (1977) formula (44), which they obtain via a tortuous route. (They reject the notion that $\Delta c/\Delta g$ measures heritability, so they redefine heritability to be that which $\Delta c/\Delta g$ does measure).

HERITABILITY ESTIMATES FROM MARTIN (1975)

Subject	Estimated h^2
English	0.79 ± 0.05
French	0.83 ± 0.07
History	0.47 ± 0.13
Geography	0.81 ± 0.06
Mathematics 1	0.81 ± 0.05
Mathematics 2	0.81 ± 0.06
Physics	0.77 ± 0.09
Chemistry	0.89 ± 0.05
Science 1	0.76 ± 0.09
Science 2	0.77 ± 0.08
IQ	0.79 ± 0.06

Obviously, equation (6) makes it easy to grind out numbers. We can apply it to the correlations in Behrman, Taubman,& Wales (1977, Table 3). We do so for three alternative values of Δg, which correspond to $g_2 = 1/3$ (approximately one of their estimates), $g_2 = 1/2$ (random mating) and $g_2 = 2/3$ (strong positive assortative mating).

UNIVARIATE HERITABILITY ESTIMATES FROM BTW DATA

	Schooling	Initial Occupation	Current Occupation	Log Earnings
Observed Δc	.22	.20	.23	.24
Estimated h^2				
Assuming Δg=2/3:	.33	.30	.34	.36
Assuming Δg=1/2:	.44	.40	.46	.48
Assuming Δg=1/3:	.66	.60	.69	.72

How sensitive are the results of the twin method to the assumption that $\Delta\rho=0$, i.e., that MZs have no more environmental resemblance than do DZs? A partial answer can be obtained from (5). We take $\Delta g = 1/2$ and, for selected values of $\Delta\rho$, plot h^2 as a function of Δc. As the figure shows, the sensitivity is substantial. For example, with $\Delta c = .2$, h^2 falls from .4 to 0 as $\Delta\rho$ rises from 0 to .2. That is, we can account for the excess observed resemblance of MZs by a (modest?) excess in environmental similarity.

Presumably we are also unsure about the value of Δg, about the systematic and the random errors in c_1 and c_2, and indeed about the absence of correlation between X and U in the first place. If enough arbitrary assumptions are made, we can identify a heavily underidentified model. Then we can grind out numbers. But why should we?

3. THE TAUBMAN VARIATION

The univariate model can be extended to allow for gene-environment correlation. In standardized form we have

$$(7) \qquad 1 = h^2 + e^2 + 2r\,he$$

$$(8) \qquad c_1 = h^2 + \rho_1 e^2 + 2r_1\,he$$

$$(9) \qquad c_2 = g_2 h^2 + \rho_2 e^2 + 2r_2\,he$$

where the new symbols r, r_1, r_2 denote the correlations of an individual's genotype with the environments of, respectively: himself, his MZ twin, and his DZ twin. That is:

$$r = \sigma_{XU}/(\sigma_X \sigma_U), \qquad r_j = \sigma_{XU*}/(\sigma_X \sigma_U), \; j = 1, 2.$$

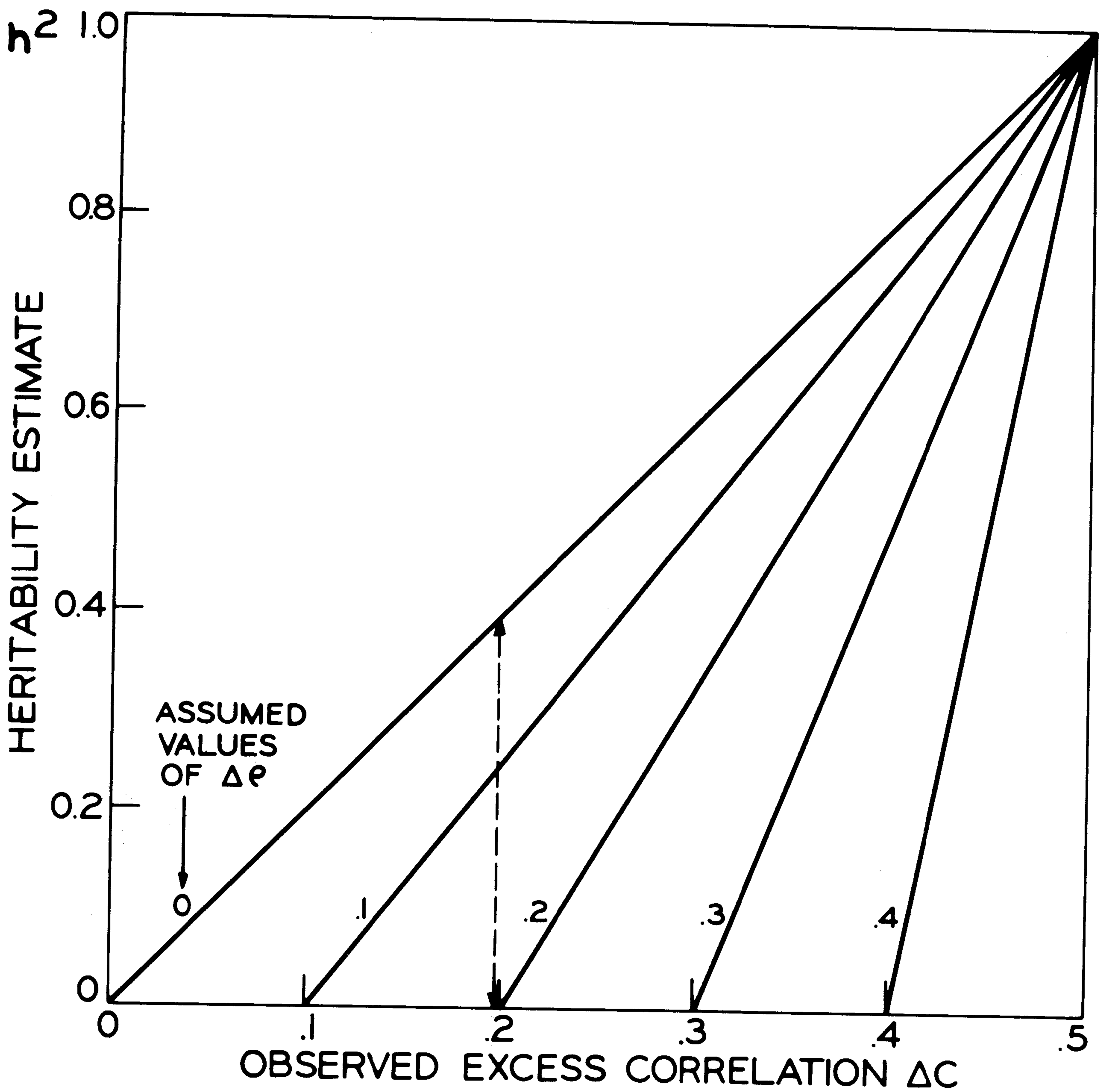
For MZs, $X = X^*$, so that $\sigma_{XU*} = \sigma_{X*U*} = \sigma_{XU}$; consequently, $r_1 = r$ automatically.

Jensen (1975) takes $r_2 = r_1 = r$; assigns alternative values to $r, \Delta\rho$, and Δg; allows for measurement error in (7); and traces out alternative estimates of h^2 and e^2, using his empirical IQ correlations $c_1 = .87$, $c_2 = .56$. Goldberger (1976) shows how sensitive his estimates are to the $r_2 = r_1$ specification. Hogarth (1974) had previously demonstrated the sensitivity to $r, \Delta\rho$, and Δg.

Taubman's (1976) model offers another variation. He takes $r_1 = r$ but $r_2 = \rho_2 r$; the latter is equivalent to $\sigma_{XU*.U} = 0$. At this point, substituting (7) into (8) and (9), we have

$$(10) \qquad c_1 = (1 - e^2) + \rho_1 e^2$$

Figure 1: Heritability Estimates from Observed Excess Correlation for Alternative Assumed Values of Environmental Excess Correlation



$$h^2 = \frac{\Delta C - \Delta \rho}{.5 - \Delta \rho}$$

$$(11) \qquad c_2 = g_2 h^2 + \rho_2 (1-h^2).$$

Here the observed MZ correlation is a weighted average of 1 and ρ_1, the weights being $1-e^2$ and e^2; the observed DZ correlation is a weighted average of g_2 and ρ_2, the weights being h^2 and $1-h^2$. If $r = 0$, then $h^2 + e^2 = 1$, and (10)-(11) will reduce to (2)-(3).

Solving (10)-(11) for h^2 and e^2 we get

$$(12) \qquad e^2 = (1-c_1) / (1-\rho_1)$$

$$(13) \qquad h^2 = (c_2 - \rho_2) / (g_2 - \rho_2).$$

Taubman (1976) takes $g_2 = 1/2$ and assumes $\rho_1 = \rho_2$ ($= \rho$, say); that is, $\Delta g = 1/2$ and $\Delta\rho = 0$. His variation thus boils down to

$$(14) \qquad e^2 = (1 - c_1) / (1 - \rho)$$

$$(15) \qquad h^2 = (c_2 - \rho) / (1/2 - \rho).$$

This model is not identified, since ρ is unknown. For twin correlations on several variables, Taubman tabulates h^2, e^2 as functions of $\rho > 0$, confining attention to the region r 0. He interprets the extreme values of h^2 obtained in this region as upper and lower bounds on heritability, and similarly for e^2.

Goldberger's (1975) analysis of (14)-(15) raised several issues. First, the upper bound on h^2 in Taubman's variation never exceeds $\Delta c / \Delta g$; this must reinforce our skepticism about (6). Second, the bounds are not confidence intervals, and give a misleading impression of precision. For example, with $c_1 = .6$ and $c_2 = .4$, Taubman's bounds give $.60 \leq e^2 \leq .66$; while if $c_1 = .7$ and $c_2 = .4$, his bounds will jump to $.40 \leq e^2 \leq .49$. For another example, if $c_1 = .6$ and $c_2 = .1$ we'd be sure that $.12 \leq h^2 \leq .20$ and $.40 \leq e^2 \leq .42$; while if $c_1 = .7$ and $c_2 = .1$, we'd be sure that the model was wrong, for no admissible solution would exist. Also if $c_2 = .5$, then the <u>only</u> admissible solution for ρ is .5 and for e^2 is $2(1 - c_1)$.[2]

4. THE JENCKS-BROWN VARIATION

Another variation is provided by Jencks & Brown (1977), who introduce gene-environment covariance, and then reparameterize to eliminate it. Their construction in effect runs as follows. Start with $Y = X + U$ where X and U have correlation r. Then

$$\sigma_{XY}/\sigma_{XX} = (\sigma_{XX} + \sigma_{XU}) / \sigma_{XX} = 1 + r\, e/h,$$

so the conditional expectation of Y given X is

$$E(Y \mid X) = (1 + r\, e/h)\; X = Z, \text{ say.}$$

The deviations from this conditional expectation are

$$Y - E(Y \mid X) = U - (r\, e/h)\; X = W, \text{ say.}$$

Then $Y = Z + W$, where "environment" W is uncorrelated with "genotype" Z. In standardized form this is

$$(16) \qquad y = h^* z + e^* w$$

where z and v are <u>uncorrelated</u> unit-variance variables, and

$$h^* = h + r e \qquad\qquad e^* = \sqrt{1 - r^2}\, e.$$

This brings us back to the model of section 2 in which genes and environment were uncorrelated. It would bring JB back there too except that they entertain the possibility that an individual's z can be correlated with his twin's w. As it happens, they eventually make a series of simplifying assumptions (including $r_1 = r_2$ and $\rho_1 = \rho_2$, which brings them to $\Delta c/\Delta g$ as their estimate of heritability, or rather of h^{*2}. Assigning Δg in turn the values .6, .5, .4 they tabulate h^{*2} for several traits from several data bases.

Note that JB have changed the traditional definition of genotype. Instead of representing the expected phenotype, for individuals of a given genetic constitution, raised over the full range of environments (or in the average environment), they take it to represent the expected phenotype, for individuals of a given genetic constitution, raised in the environments in which they are presently found (cf. Falconer (1960, pp. 132-133)). This is one way to eliminate bias: just announce that the parameter of interest is the expectation of your estimator. It is unclear why h^{*2} is an interesting parameter: to what question is it the answer? JB hardly help matters by referring to h^{*2} as "broad heritability."

5. A MULTIVARIATE MODEL

We now consider a model for several observed variables. An individual's observed vector of phenotypes is determined as a linear form in three unobserved vectors representing genotypes, environments, and errors. In standardized form we write

$$\underline{y} = H\underline{x} + E\underline{u} + F\underline{v}$$

where

$\underline{y}$ (m x 1) is the phenotypic vector

$\underline{x}$ (p x 1) is the genotypic vector

$\underline{u}$ (q x 1) is the environmental vector

$\underline{v}$ (m x 1) is the error vector

and

H (m x p) is the heritability-loading matrix

E (m x q) is the environmentability-loading matrix

F (m x m) is the <u>diagonal</u> error-loading matrix.

It is assumed that all elements of $\underline{x}$, $\underline{u}$, $\underline{v}$ are uncorrelated, so that the m x m phenotypic correlation (= covariance) matrix is

(17) $$S = H\ H' + E\ E' + F^2.$$

The individual is paired with another for whom $\underline{y}^* = H\,\underline{x}^* + E\,\underline{u}^* + F\,\underline{v}^*$. Their m x m phenotypic cross-correlation (= cross-covariance) matrix is

$$C = H\ G\ H' + E\ R\ E'$$

where

$G = \Sigma_{xx^*}$ (p x p) is the <u>diagonal</u> genotypic cross-correlation matrix

$R = \Sigma_{uu^*}$ (q x q) is the <u>diagonal</u> environmental cross-correlation matrix,

on the assumption that the only nonzero cross-correlations are those between corresponding elements of $\underline{x}$ and $\underline{x}^*$, and those between corresponding elements of $\underline{u}$ and $\underline{u}^*$.

Introduce the subscripts 1 for MZ and 2 for DZ. We now have

(18) $$C_1 = H\ H' + E\ R_1\ E'$$

(19) $$C_2 = H\ G_2 H' + E\ R_2\ E'$$

since $\underline{x}^* = \underline{x}$ for MZs so that $G_1 = I$. Thus

$$C_1 - C_2 = H\,(I - G_2)\ H' + E\ (R_1 - R_2)\ E',$$

or, more compactly,

(20) $$\Delta C = H(\Delta G)H' + E(\Delta R)E'.$$

The multivariate formulas (18)-(20) resemble the univariate formulas (2)-(4). If $p + q < m$, then ΔC is short-ranked, which is a testable restriction. But to identify parameters, further assumptions are still required.

Assume that the environmental correlations for MZ s are the same as for DZ s, that is $\Delta R = 0$. Then

(21) $$\Delta C = H(\Delta G)H',$$

i.e., all the excess observed resemblance of MZs is attributable to their excess genetic resemblance. If $p < m$, then ΔC in (21) is short-ranked, which is a testable restriction. This seems to be the specification of Bock & Vandenberg (1968, pp. 246-248), who fit $m = 8$ test scores with $p = 3$ genetic factors, using samples of about 100 pairs of MZs and 80 pairs of DZs. A similar approach was applied to another data set by Bock (1973).[3]

When the objective is to determine the number of distinct genetic factors, there is no need to further specify the structure of ΔG. But the procedure may be pushed further by assuming that the diagonal elements of G_2 are all equal. Then ΔG reduces to $\Delta G = (\Delta g)I$, say; and

(22) $$\Delta C = (\Delta g)HH' = H^* H^{*\prime},$$

say, where $H^* = \sqrt{\Delta g}\,H$. With adequate normalizations--e.g. triangularity or orthogonality of H and hence of H*--one may then in principle determine H* from ΔC. If the scalar Δg is also given, H itself is determined. Observing that the diagonal elements of S are unities, one may take the diagonal elements of H H' as estimates of the heritabilities of the respective traits, or for that matter, take $\mathrm{tr}(HH')/\mathrm{tr}(S) = \mathrm{tr}(HH')/m$ as a scalar measure of heritability for the set of phenotypes. Furthermore, the ratio of an off-diagonal element of HH' to the square root of the product of the corresponding diagonal elements of HH', could be interpreted as the correlation between the genetic factors which impinge on a pair of phenotypes.[4]

But why do all this when $\Delta R = 0$ was an arbitrary assumption in the first place? Skepticism about the multivariate twin method seems to be justified. On the other hand, it may well be interesting to study the dimensionality of ΔC without dramatizing the results in terms of heredity vs. environment.

6. THE BEHRMAN-TAUBMAN-WALES VARIATIONS

An elaboration of the multivariate model is employed by Behrman, Taubman, & Wales (1977), who analyze the schooling, initial occupation, current occupation, and log earnings of 1019 MZ pairs and 907 DZ pairs, all white males. The reduced form of the typical BTW structural model, in standardized form, is

(23) $$\underline{y} = H\,\underline{x} + \underline{e}\,u + F\,\underline{v},$$

where, with m = 4, p = 4, q = 1:

$\underline{y}$ is 4 x 1, $\underline{x}$ is 4 x 1, u is 1 x 1, $\underline{v}$ is 4 x 1

H is the 4 x 4 triangular heritability-loading matrix

$\underline{e}$ is the 4 x 1 environmentability-loading vector

F is the 4 x 4 triangular coefficient matrix.

Note that F is no longer diagonal, since the reduced-form disturbances are permitted to be correlated. Indeed F^{-1} is the coefficient matrix of their recursive structural model for the observable variables.

Assuming that all elements of $\underline{x}$, u, and $\underline{v}$ are uncorrelated, the phenotypic correlation matrix is

(24) $$S = HH' + \underline{ee}' + FF'.$$

The individual is paired with another for whom $\underline{y}^* = H\,\underline{x}^* + \underline{e}\,u^* + F\,\underline{v}^*$. Their phenotypic cross-correlation matrix is

$$C = gHH' + \rho\underline{ee}'$$

where

g = genotypic cross-correlation

ρ = environmental cross-correlation,

on the assumptions that the only nonzero cross-correlations are those between corresponding elements of $\underline{x}$ and $\underline{x}^*$ (all being equal) and that between u and u*. For MZs, subscripted 1, and DZs, subscripted 2, we have

(25) $$C_1 = HH' + \rho_1\,\underline{ee}'$$

(26) $$C_2 = g_2HH' + \rho_2\,\underline{ee}',$$

since $g_1 = 1$.

Thus

(27) $$C_1 - C_2 = (1-g_2)\,HH' + (\rho_1 - \rho_2)\,\underline{ee}'.$$

Also, multiplying (25) by g_2 and subtracting from (26):

(28) $$C_2 - g_2\,C_1 = (\rho_2 - g_2\,\rho_1)\,\underline{ee}'.$$

Subtracting (25) from (24) we also have

$$S - C_1 = (1 - \rho_1)\,\underline{e}\underline{e}' + FF'. \tag{29}$$

Equations (27)-(29) contain the same information as (24)-(26). Without further restrictions, the parameters are not identified.

Now turn to the series of models in Sections VIII-IX of the BTW article, which we label by their table numbers. The main objective of the BTW study is presumably the estimation of the structural coefficient matrix F^{-1}. They also emphasize the estimated allocation of variance among genetic, environmental, and disturbance factors: in our notation this allocation is given by the diagonal elements of HH', $\underline{e}\underline{e}'$, and FF'. We critically examine how the various parameters are determined from S, C_1, and C_2.[5]

Model 5. It is assumed that $g_2 = 1/2$ and that $\rho_1 = \rho_2 = 1$. So (27) gives $C_1 - C_2 = 1/2\ HH'$ which determines the triangular matrix H, and (28) gives $C_2 - 1/2\ C_1 = 1/2\ \underline{e}\underline{e}'$ which determines the vector $\underline{e}$. Finally (29) gives $S - C_1 = FF'$ which determines the triangular matrix F. (Throughout we use "determines" as shorthand for "determines uniquely apart from irrelevant signs.")

Model 5a. Here it is assumed that $g_2 = 1/2$ and $\rho_1 = \rho_2$, but their common value (ρ, say) is not fixed a priori. We have H from (27) as before, and (28) gives

$$C_2 - 1/2\ C_1 = 1/2\ \rho\ \underline{e}\underline{e}' = \underline{q}\,\underline{q}',$$

say, where $\underline{q} = \sqrt{\rho/2}\ \underline{e}$. That determines $\underline{q}$. Then (29) leaves us with

$$S - C_1 = (2(1-\rho)/\rho)\,\underline{q}\,\underline{q}' + FF'. \tag{30}$$

Evidently neither ρ nor F are yet determined. An additional restriction is required. What BTW assume is that $f^{42} = 0$. That suffices to determine $2(1-\rho)/\rho$, hence ρ and $\underline{e}$, and (the remaining unknown elements of) F in (30).

What $f^{42} = 0$ says is that there is no direct path from initial occupation to log earnings. This restriction, introduced with little explanation at the beginning of Section II, is employed throughout BTW. Without it, Model 5a—and some of their other models—would not be identified. But Model 5 is identified without that restriction, so it could have been tested there. It's unclear why BTW failed to do so, nor is it clear why they excluded initial occupation from even the log-earnings regressions reported in their Tables 4 and 4a. From the observed correlations given in their Table 3, we can compute the (standardized) multiple regression of log-earnings on all three prior observed variables. Our results are tabulated below.

STANDARDIZED LOG-EARNINGS REGRESSIONS IN BTW DATA

Dependent variable: Log earnings

	Independent variables		
Data	Schooling	Initial Occ.	Current Occ.
MZ Individuals	.30	.13	.13
DZ Individuals	.29	.15	.14
Within MZs	.11	.02	.12
Within DZs	.23	.12	.10

The beta-weight is as large for initial occupation as for current occupation, within-DZs as well as across individuals. The case for $f^{42} = 0$ may rest on the within-MZ regressions.

Model 5b. Here it is assumed that $g_2 = 1/2$, that $\rho_2 = .8\,\rho_1$, while $\rho_1 = \rho$, say, is free. Equation (28) gives $C_2 - 1/2\,C_1 = .3\,\rho\,\underline{e}\underline{e}' = \underline{q}\,\underline{q}'$, say, where $\underline{q} = \sqrt{.3\rho}\,\underline{e}$. That determines $\underline{q}$. Equation (27) gives

$$C_1 - C_2 = 1/2\,HH' + 2/3\,\underline{q}\,\underline{q}',$$

which determines the triangular matrix H. Finally, (29) gives

$$S - C_1 = ((1-\rho)\,/\,.3\rho)\ \underline{q}\,\underline{q}' + FF'.$$

The restriction $f^{42} = 0$ determines $(1 - \rho)/.3\rho$, hence ρ and $\underline{e}$, and F.

Here, as in Taubman (1976), the $\Delta\rho = 0$ assumption is replaced by a fixed ratio of ρ_2 to ρ_1. Now BTW point out that the fit of model 5b is the same as that of 5a, and that the estimates of disturbance variances and structural coefficients are unchanged. What do change are the estimated loadings on the latent variables, and consequently the allocation of phenotypic variances into genetic and environmental components. Evidently $\rho_2 = \rho_1$ and $\rho_2 = .8\rho_1$ are observationally equivalent.[6]

To clarify this point, refer back to (24)-(26). Let H, $\underline{e}$, F, $g_2 = 1/2$, ρ_1, ρ_2 denote the true parameter values. Let λ be an arbitrary positive scalar. Then let

$$\underline{e}^* = (1/\sqrt{\lambda})\,\underline{e}$$

$$\rho_1^* = 1 - \lambda\,(1 - \rho_1)$$

$$\rho_2^* = 1/2 + (\lambda_2 - 1/2)$$

$$M = HH' + (1 - \frac{1}{\lambda}) \underline{ee'}$$

H^* = triangular matrix such that $H^*H^{*'} = M$.

Replacing the true values in (24)-(26) by their starred counterparts, while retaining F and $g_2 = 1/2$, and collecting terms, one finds that the equations are preserved. Thus the pseudo-values are observationally indistinguishable from the true values. And the choice of λ is limited only by the requirements that the ρs remain in the unit interval and that M remain nonnegative definite. In 5a, the indeterminancy was resolved by equating the two environmental correlations; in 5b it was resolved by setting the DZ environmental correlation at eight-tenths of the MZ environmental correlation.

The estimates in 5b are simple translations of those in 5a. In particular, we can move from the $\rho_1 = \rho_2 = .86$ estimates in 5a to the $\rho_1^* = .91$, $\rho_2^* = .728$ estimates in 5b by choosing $\lambda = .64$. For that matter, we can move from the $\rho_1 = \rho_2 = .86$ solution to a $\rho_1^* = .94$, $\rho_2^* = .66$ solution (call it Model 50) by choosing $\lambda = .44$. Doing so leaves the goodness of fit, the disturbance variance estimates, and the structural coefficient estimates untouched. It does change the loadings on the latent variables, and consequently the allocations of variance. We calculate the latter and tabulate them below.

ALTERNATIVE VARIANCE ALLOCATIONS FOR BTW DATA

	Model 5a				Model 5b				Model 50			
Genetics	46	33	31	47	26	21	25	43	1	7	17	36
Environment	35	21	11	9	55	33	17	13	80	47	25	20
Disturbance	19	46	58	44	19	46	58	44	19	46	58	44
Total	100	100	100	100	100	100	100	100	100	100	100	100

In each panel the four columns refer respectively to Schooling, Initial Occupation, Current Occupation, and Log Earnings. The entries are percentages.

Observing the sensitivity of the variance allocations to the changes in $\Delta\rho$, for observationally equivalent models, hardly restores one's faith in heritability estimates produced by the multivariate twin methods.

Model 5c. Here g_2 is free and it is assumed that $\rho_1 = \rho_2 = 1$. From (29) we get $S - C_1 = FF'$ which determines F, even without the restriction $f^{42} = 0$. Equation (28) gives us

$$C_2 - g_2 C_1 = (1 - g_2) \underline{e}\underline{e}',$$

which determines g_2 as the scalar which makes $C_2 - g_2 C_1$ a rank-one matrix. Along with it, we get $\sqrt{1-g_2}\,\underline{e}$, and thus $\underline{e}$ itself. Finally (27) gives $C_1 - C_2 = (1 - g_2) HH'$; with g_2 in hand, this determines H.

Model 5d. Here g_2 is free, and $\rho_1 = \rho_2$ but their common value ρ is not fixed a priori. From (28) we have

$$C_2 - g_2 C_1 = \rho (1 - g_2) \underline{e}\underline{e}',$$

which determines g_2 as the scalar which makes $C_2 - g_2 C_1$ a rank-one matrix. Along with it we get $\sqrt{1-g_2}\,\underline{q}$, where now $\underline{q} = \sqrt{\rho}\,\underline{e}$, and thus $\underline{q}$ is determined. With g_2 in hand, (27) determines H as before. Finally (29) gives $S - C_1 = ((1 - \rho)/\rho)\, \underline{q}\,\underline{q}' + FF'$.

The restriction $f^{42} = 0$ determines $(1-\rho)/\rho$, hence ρ and $\underline{e}$, and F.

For 5d as well as 5c, BTW estimate g_2 to be .34. They rationalize this low value rather casually in terms of negative assortative mating and dominance. But they do not consider the extent of negative assortative mating and/or dominance which are needed to drive g_2 down so far from its random-mating—no-dominance value of .50. A little arithmetic suggests that quite exotic values are required. In the classicial genetic model, the genotypic correlation between DZs may be written as

(31) $$g_2 = 1/2\, n\, (1 + mnh^2) + 1/4\, (1-n)$$

where

h^2 = ratio of total genetic variance to total phenotypic variance

n = ratio of additive genetic variance to total genetic variance

m = phenotypic correlation of parents.

(Cf. Burt & Howard (1956, pp. 113-116)). With random mating (m = 0) and no dominance (n = 1), this gives $g_2 = .50$. The BTW estimates are approximately $g_2 = .35$ and $h^2 = .35$. Inserting these values into (31) gives

(32) $$m = (4 - 10n)/(7n^2).$$

Unless they believe in very strong negative assortative mating, BTW must believe in very weak additive genetic effects: from (32), $m > -1/2$ requires $n < .48$.[7]

A skeptical interpretation of the free g_2 parameter is that it is proxying for a difference in environmental correlation which was assumed away in these models by forcing $\Delta\rho = 0$.

Models 5e, 5f. Discussion of these two models is omitted here because it would require extending our framework. In any event, both models impose $\Delta\rho = 0$.

Model 5g. This, too, requires extending the framework, but we discuss it because it represents BTW's only version of a pure-environmentalist model. The genetic components are dropped, and are replaced by three additional environmental components. The reduced form for an individual is

$$\underline{y} = E\,\underline{u} + F\,\underline{v}$$

where

$\underline{y}$ is 4 x 1, $\underline{u}$ is 4 x 1, $\underline{v}$ is 4 x 1

E is 4 x 4 triangular

F is 4 x 4 triangular (with $f^{42} = 0$).

The individual is paired with another, for whom $\underline{y}^* = E\underline{u}^* + F\underline{v}^*$. It is assumed that the only nonzero correlations among latent variables are those between corresponding elements of $\underline{u}$ and $\underline{u}^*$ (all equal to ρ_1 for MZs, all equal to ρ_2 for DZs). Thus the phenotypic correlation matrices are

$$S = EE' + FF', \qquad C_1 = \rho_1\,EE', \qquad C_2 = \rho_2\,EE'.$$

Identification is verified by noting that the triangular matrix E is determined up to a scalar multiple from either C_1 or C_2; the scalar is then determined along with F from S and the restriction $f^{42} = 0$. Returning to C_1 and C_2 then determines ρ_1 and ρ_2.

We note that the ratio ρ_2/ρ_1 is determinable without the restriction $f^{42} = 0$. Indeed a key implication of this model is that DZ phenotypic correlations are proportional to those for MZs: $C_2 = (\rho_2/\rho_1)\,C_1$. (Of course, if the four elements of $\underline{u}$ were permitted to have different cross-correlations, this implication would vanish.)

BTW report that 5g fits worse than the preceding models: "the log of the likelihood function is smaller than our previous best estimates by 16." They neglect to mention that 5g uses at least 3 fewer parameters. An instructive way to examine the fit is to compare the predicted (i.e., "reproduced") correlation matrices with the observed ones. We tabulate

COMPARISON OF TWO MODELS FOR BTW DATA

Correlation Matrices

	Observed				Predicted Model 5c				Predicted Model 5g			
Individuals*												
	1.00	.53	.53	.44	1.00	.53	.52	.44	1.00	.52	.52	.44
		1.00	.44	.36		1.00	.44	.35		1.00	.44	.35
			1.00	.35			1.00	.35			1.00	.35
				1.00				1.00				1.00
Cross-MZ												
	.76	.47	.44	.40	.77	.47	.43	.41	.76	.47	.42	.40
		.53	.35	.32		.53	.34	.33		.51	.33	.32
			.43	.27			.41	.27			.39	.27
				.54				.57				.53
Cross-DZ												
	.54	.37	.29	.29	.53	.36	.30	.29	.50	.31	.28	.26
		.33	.22	.22		.33	.23	.22		.33	.22	.21
			.20	.19			.23	.17			.25	.17
				.30				.27				.34
Variance Allocation (percentages)												
Genetics					36	30	28	45	0	0	0	0
Environment					41	22	13	12	81	54	41	56
Disturbance					23	48	59	43	19	46	59	44
Total					100	100	100	100	100	100	100	100

* For individuals, the observed figures are averages of correlations for MZ individuals & DZ individuals. In each panel, the four columns refer respectively to Schooling, Initial Occupation, Current Occupation, and Log Earnings.

the results for 5c and 5g; the underlined entries are those which differ by .03 or more from their observed values. This informal comparison indicates that 5c does fit better than 5g, and that the superiority is not overwhelming.

At the bottom of our table are the variance allocations implied by the two models. Recall that in 5c $\rho_1 = \rho_2 = 1$; in 5g BTW estimate $\rho_1 = .94$, $\rho_2 = .61$. Broadly speaking, the heritability estimates fall from about .33 to 0 as $\Delta\rho$ is allowed to rise from 0 (in 5c) to .33 (in 5g). This is strikingly reminiscent of our arithmetic in the univariate model of Section 2. Again the sensitivity of heritability estimates to assumptions about $\Delta\rho$ continues to justify skepticism. In the present case, to be sure, the elimination of heritability is accompanied by a worsening of the fit.

BTW, at the end of Section IX, report that they also added a single genetic factor (with $g_2 = 1/2$) to Model 5g. That new model gave "approximately the same" likelihood as their earlier "genetic" ones. They do not tabulate the estimates, so we are left to wonder whether "approximately the same" means "slightly better" or "slightly worse". And we are left to wonder about the variance allocations in the new model. In any event, a skeptical interpretation of the new factor is that it is proxying for an environmental factor whose cross-correlations differ from ρ_1, ρ_2. (Recall that in 5g all four environmental factors are forced to have the same ρ_1, ρ_2.)

7. AN AGNOSTIC PROPOSAL

It appears that heritability estimates produced by the twin methods are quite dependent upon assumptions about differential environmental correlations. We are left with the data which show MZs more highly correlated on observed variables than DZs, but have no basis for allocating these differences between genetic and environmental factors.

As an agnostic alternative to the BTW heredity-environment specifications we propose the following reduced form for an individual:

$$\underline{y} = A\,\underline{z} + F\,\underline{v}$$

where

$\underline{y}$ (m x 1) is the observed vector

$\underline{z}$ (m x 1) is the latent-variable vector

$\underline{v}$ (m x 1) is the structural disturbance vector

and

A (m x m) is <u>triangular</u>

F (m x m) is <u>triangular</u>.

All elements of $\underline{z}$ and $\underline{v}$ are uncorrelated, so that the phenotypic correlation matrix is

$$S = AA' + FF'.$$

The individual is paired with another for whom $\underline{y}^* = A\,\underline{z}^* + F\,\underline{v}^*$. Their observed cross-correlation matrix is

$$C = AQA'$$

where Q is the diagonal latent-variable cross-correlation matrix, on the assumption that the only nonzero cross-correlations are those between corresponding elements of $\underline{z}$ and $\underline{z}^*$. The diagonal elements of Q are not necessarily equal to one another. Furthermore they differ as between the two twin types. For MZs and DZs respectively we have

$$C_1 = AQ_1A' , \qquad C_2 = AQ_2A' .$$

Thus $\Delta C = A(\Delta Q)A'$: the observed differences between MZs and DZs are accounted for by the differences between their correlations on the latent variables.

Since F and A are unrestricted apart from triangularity, S is essentially uninformative about the parameters. The levels of Q_1 and Q_2 are also indeterminate, so we reparameterize by normalizing on MZs rather than on individuals:

$$S = A^*Q^oA^{*\prime} + FF'$$

$$C_1 = A^*A^{*\prime}$$

$$C_2 = A^*Q^*A^{*\prime}$$

where

$$A^* = AQ_1^{\frac{1}{2}} , \quad Q^o = Q_1^{-1} , \quad Q^* = Q_1^{-\frac{1}{2}}\,Q_2\,Q_1^{-\frac{1}{2}} .$$

Fitting A* (triangular) and Q* (diagonal) to the BTW cross-correlation matrices C_1 and C_2 gives the AG Model results in the following table; again the underlined entries are those which differ by .03 or more from their observed values. We don't bother to extract F from S since the elements of Q (diagonal) can be chosen fairly arbitrarily, so that F is fairly indeterminate. The AG Model uses 14 parameters to fit the 20 cross-correlations. In terms of predicted correlations, the fit is somewhat worse than the BTW models. The estimated diagonal elements of Q* simply indicate that on the latent variables DZs are less highly correlated than MZs, the differences being different for the several latent variables.

In the lower right-hand corner of the table are our A*-estimates translated into natural units of the observable variables. The columns

AGNOSTIC MODEL FOR BTW DATA

Correlation Matrices

Cross-MZ

Observed				Predicted AG Model			
.76	.47	.47	.40	.72	.51	.42	.40
	.53	.35	.32		.58	.35	.34
		.43	.27			.40	.30
			.54				.57

Cross-DZ

Observed				Predicted AG Model			
.54	.37	.29	.29	.54	.36	.30	.29
	.33	.22	.22		.32	.22	.21
		.20	.19			.21	.18
			.30				.29

Parameter Estimates

A* (standardized form)				A* (destandardized)			
.87	0	0	0	2.68	0	0	0
.58	.49	0	0	1.42	1.20	0	0
.48	.14	.39	0	1.02	.31	.83	0
.46	.14	.14	.56	.26	.08	.08	.31

Q* (diagonal elements)

.71	.32	.25	.39

here bear a striking resemblance to those in BTW Tables 5, 5a-5g, especially if one views our first latent variable as a weighted sum of their G and N variables.

8. CONCLUSION

Heritability estimates produced by the twin methods are shaky. Is this a cause for concern? The answer would be yes if the estimates served some useful purpose. Do they?

Suppose that unmeasured family background variables play an important role in the socioeconomic achievement process. To get unbiased estimates of the effects of measured variables, we would want to control for the unmeasured variables. Suppose further that the unmeasured variables are more highly correlated for MZs than for DZs. Then within-MZ regressions might provide the most effective control. But there is no need to sort out the background into genetic and environmental components.

Then is the allocation needed for some other purpose? The rationale offered by BTW is:

> We think that a portion of the equity criteria within and between generations is related to the extent to which a person's earnings are due to his own efforts versus those of the parents who bore and reared him.

But here again the distinction between genes and environment (between bearing and rearing) is irrelevant.[8]

BTW go on to say that

> It is of some interest to see how unequal the [income] distribution would be if just family environments were equalized.

And towards the end of their article, BTW state that their substantial estimates of heritability

> imply that even extreme policies to assure equality of opportunity by eliminating all differences in environments (including those due to the family) would not eliminate much of the family contribution to the welfare of offspring.

It is surprising to find economists using current variance allocations to predict the outcomes of extreme policy changes.

Then why do quantitative geneticists measure heritability? As far as I know, they do so not because heritability is informative about the outcomes of environmental changes, but because it is informative about the outcomes of selective breeding programs under constant environmental conditions. If this reading is correct, economists and sociologists would do well to abandon heritability estimation, and accept Jencks et al.'s (1972, p. 76) remark

> Indeed our main conclusion after some years of work on this problem is that mathematical estimates of heritability tell us almost nothing about anything important.

FOOTNOTES

[1]From an informal sketch of the twin method given by P. Mittler (1971, Chapter 3), we select the following passages:

> Reduced to its simplest form, the twin method assumes that any difference within identical pairs must be due to environmental or at least non-genetic causes, whereas differences within fraternal pairs are due to both environmental and genetic factors. The extent to which identical twins resemble each other more than fraternal twins is held to reflect the strength of the genetic contribution to a characteristic. . .
>
> (p. 45)

> The classical twin study method is open to criticism for confusing the genuinely greater physical and genetic similarity of MZ twins with the greater environmental similarity to which they might be exposed as a result of parents, siblings and peers treating them in the same way. Any environmentally created similarities, would, of course, tend to increase the similarities between identical twins, and lend spurious support to a genetic interpretation of the data. . .
>
> (p. 52)

> . . .[I]t is probably reasonable to conclude that there is little empirical justification for the basic assumption of the twin method to the effect that the environmental contribution to the within-pair variance of a characteristic is the same for identical and fraternal twins. . .
>
> (pp. 52-53)

[2]Bounding procedures for underidentified models have been used by others, e.g. Chamberlain & Griliches (1975). Perhaps similar issues should

be raised there.

[3]The origins of the multivariate twin method date back earlier. For example, Kempthorne & Osborne (1961) wrote:

> Just as we can develop an analysis of variance for one measurement, we can develop an analysis of covariance for two measurements. . . The same general considerations will then be found to hold as for components of variance. . .
>
> (pp. 330-333)

Among those considerations were

> . . .In theory at least, an efficient method for appraising the heredity-environment problem in man, particularly with respect to complex or quantitative inheritance, is by the study of twins. . . This formulation. . . assumes. . . .that the environmental forces which affect within-pair character differences are comparable within the two types of twins.
>
> (pp. 320-321)

> . . . The same adolescent environment will be largely shared by monozygotic twins, probably to a lesser extent by like-sexed dizygotic twins, and may be quite different for unlike-sexed pairs. . . Consequently, assumptions as to shared environment become less certain.
>
> (p. 325)

They went on to write that

> The existence of many possible explanations with the limited data at our disposal is a part of our general theme. . .One may expect that the closer the genetic relationships, the closer will be the environmental covariance. If, of course, both changed in the same proportion, separation of genetic and environmental causation will be difficult. . .
>
> (pp. 336-337)

More recent examples of the multivariate twin method can be found in Eaves & Gale (1974).

[4]This approach is, in effect, used by Jencks & Brown (1977) in their analysis of education and test score. With p = 2 = m, the correlation between the genetic determinants of the two phenotypes is, from (22):

$$\Delta c_{12}/\sqrt{\Delta c_{11} \Delta c_{22}}$$

where the subscripts now label observed variables. This is formula (84) derived by JB. It can also be found in the discussion on p. 300 of Vandenberg (1965). See also Falconer (1960, pp. 311-318).

[5]Our identification analysis is closely related to that of Chamberlain (1977).

[6]Readers of Taubman (1976), Behrman & Taubman (1976), and Taubman & Wales (1976) may have thought that assumptions on ρ_2 vs. ρ_1 were testable.

[7]Furthermore, negative assortative mating and weak additive genetic effects dampen the intergenerational implications of high heritability: in the classical genetic model, the genotypic correlation between parent and child is given by

$$g_o = 1/2\, n\, (1 + m).$$

[8]Nor does the accompanying footnote help:

> If some people are born or reared to be lazy, this criteria suggests that they are still entitled to transfer payments.

REFERENCES

J. Behrman & P. Taubman (1976), "Intergenerational transmission of income and wealth," American Economic Review, Vol. 66, May 1976, pp. 436-440.

J. Behrman, P. Taubman, & T. Wales (1977), "Controlling for and measuring the effects of genetics and family environment in equations for schooling and labor market success," this volume.

R. D. Bock (1973), "Word and image: sources of the verbal and spatial factors in mental test scores," Psychometrika, Vol. 38, December 1973, pp. 437-457.

R. D. Bock and S. G. Vandenberg (1968), "Components of heritable variation in mental test scores," Chapter 14, pp. 233-260 in S. G. Vandenberg, editor, Progress in Human Behavior Genetics, Baltimore: Johns Hopkins.

C. Burt & M. Howard (1956), "The multifactorial theory of inheritance and its application to intelligence," British Journal of Statistical Psychology, Vol. 8, May 1956, pp. 95-131.

G. Chamberlain (1977), "Are brothers as good as twins?", this volume.

G. Chamberlain & Z. Griliches (1975), "Unobservables with a variance-components structure: ability, schooling and the economic success of brothers," International Economic Review, Vol. 16, June 1975, pp. 422-450.

L. J. Eaves & J. S. Gale (1974), "A method for analyzing the genetic basis of covariation," Behavior Genetics, Vol. 4, 1974, pp. 253-267.

D. S. Falconer (1960), Introduction to Quantitative Genetics, Edinburgh & London: Oliver & Boyd.

A. S. Goldberger (1975), "On Taubman's model for twins," manuscript, June 1975, 19 pp.

A. S. Goldberger (1976), "On Jensen's method for twins," Educational Psychologist, Vol. 12, 1976, pp. 79-82.

R. M. Hogarth (1974), "Monozygotic and dizygotic twins reared together: sensitivity of heritability estimates," British Journal of Mathematical & Statistical Psychology, Vol. 27, May 1974, pp. 1-13.

C. Jencks et al. (1972), Inequality, New York: Basic Books.

C. Jencks & M. Brown (1977), "Genes and social stratification: a methodological exploration with illustrative data," this volume.

A. R. Jensen (1967), "Estimation of limits of heritability of traits by comparison of monozygotic and dizygotic twins," Proceedings of the National Academy of Science, Vol. 58, 1967, pp. 149-156.

A. R. Jensen (1975), "The meaning of heritability in the behavioral sciences," Educational Psychologist, Vol. 11, 1975, pp. 171-183.

O. Kempthorne & R. H. Osborne (1961), "The interpretation of twin data," American Journal of Human Genetics, Vol. 13, 1961, pp. 320-339.

N. G. Martin (1975), "The inheritance of scholastic abilities in a sample of twins: II. Genetical analysis of examination results," Annals of Human Genetics, Vol. 39, pp. 219-229.

P. Mittler (1971), The Study of Twins, Harmondsworth, Middlesex: Penguin Books.

P. Taubman (1976), "The determinants of earnings: genetics, family and other environments a study of white male twins," American Economic Review, Vol. 66, December 1976.

P. Taubman & T. Wales (1976), "The path to labor market success: the role of unobservable genetic and environmental factors," mimeographed, 35 pp.

S. G. Vandenberg (1965), editor, Methods and Goals in Human Behavior Genetics, New York: Academic Press.