The Determinants of Earnings: Genetics, Family, and Other Environments; A Study of White Male Twins

By PAUL TAUBMAN*

Two major concepts that economists employ when discussing the distribution and redistribution of income are equity and efficiency.[1] While there is no standard measure of equity, the term is generally defined by the extent of the inequality and oftentimes by the source of the inequality. For example, even neoclassical economists describe discrimination as "bad," and the word "nepotism" connotes unfairness.[2]

Some economists, including myself, are more tolerant of inequality arising from an individual's hard work and effort than that arising from contributions of one's parents. Parents can aid or handicap their offspring in diverse ways. Parents can transfer assets to or purchase education for their children. Parents may also affect their child's earning potential through diet and inculcation of values. These are two examples of family environmental effects. Biological parents also endow a child through the genes they contribute. The total of family environmental and genetic effects will be labeled the family effect.

For some equity related questions, it may not be necessary to break up the family effects into components. But there are other questions that are related in part to equity for which it seems appropriate to distinguish genetic and family environmental effects.

For example, while a particular redistributive mechanism such as schooling may be effective in offsetting both "poor" genes and "poor" home environment, other mechanisms may be more effective against one or the other.[3] Moreover, both income transfer and compensatory training programs must be paid for from tax programs which will induce disincentives and losses in efficiency. But the tradeoffs with efficiency may be different if the source of the inequality is genetic or environmental. In addition, one important aspect of equity is the degree of social mobility. In some sense the greater the variability in parent/child mobility among siblings, the greater is the intergenerational mobility. The variation among siblings in family effects may depend on the relative importance of genetics and

---

* Professor of economics, University of Pennsylvania. I would like to thank J. Rotwitt and H. Rosenthal, who uncovered the existence of the data set on which this work is based as part of their research in a senior seminar. I would also like to thank the Medical Followup Agency of the National Academy of Science-National Research Council (*NAS-NRC*) for making the data available and for conducting the surveys, and especially James Norman, A. Hiram Simon, and Hatsumi Hamamura for their assistance in the analysis and survey. Jere Behrman, Arthur Goldberger, Zvi Griliches, Abraham Haspel, and Christopher Jencks have given me many helpful suggestions on this topic and corrected numerous errors. Barbara Atrostic and Shahrokh Fardoust supplied able research assistance. The research in this paper was supported by *NSF* Grant SOC 73-05543. An earlier version of this paper was presented at a Conference on Inequality at the Institution for Advanced Studies, which was supported by the Sloane Foundation.

[1] See for example Theodore Schultz.

[2] See for example Kenneth Arrow.

[3] For example, estate taxation may be effective against nepotism while certain medicines may be effective against diseases caused by a particular genetic defect.

family environment. Finally, the distribution of some aspects of family environment, for example, income, has changed by age cohort in this century in the United States. Cross-cohort comparisons will be strengthened if we can distinguish those sources of inequality that have and have not been stable.

In this paper I introduce a major new sample, all of whose members are twins. I then use a variance components model to obtain some estimates of the contributions of genetics, family, and other environments to the variance of the *log* of earnings of white males around age 50. The estimate for other environment can be obtained in a rather straightforward way, while a number of strong assumptions must be made to separate the contributions of genetics and family environment.

Section I of this paper presents some information about twins and necessary definitions. The next section develops the statistical model and discusses the validity of some underlying assumptions. Section III will introduce the National Academy of Science-National Research Council (*NAS-NRC*) twin sample to economists. Section IV applies statistical techniques to partition the variance of earnings of white males about 50 years old into that due to variances in genetic endowments, common (family) environment, noncommon environment, and (twice) the covariance of environment and genetic endowments.

I. Genetics, Environment, and Twin Type

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.

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 the genetic information of the original fertilized egg (barring mutations).[5]

II. The Model

In this paper I draw heavily on previous work in quantitative genetics that flows from an initial classic paper by R. A. Fisher, and that yields a form of a variance component model.[6] In a separate paper (1976) I use twins as controls in regression analysis. In that paper I demonstrate that standard proxies for environment (and, to a lesser extent, genetics) such as parental education and occupation are statistically significant determinants of earnings. However, I also show that these variables are far from perfect proxies. In this paper I do not use those proxies, but instead infer the results from statistical techniques.

Our economic model is that a person is paid a wage (net of cost of investments in on-the-job training) equal to his marginal product. It is further assumed that marginal product is determined by skills which are produced by combining various genetic

---

[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 L. Cavalli-Sforza and W. Bodmer.

[5] For a more thorough and rigorous treatment of the genetic and biological underpinnings, see Cavalli-Sforza and Bodmer.

[6] More accessible versions are in Oscar Kempthorne, D. Falconer, and J. Jinks and D. Fulker.

endowments with a variety of environments.[7] Suppose the substitution of the production functions into the marginal product condition yields the following hedonic price index for earnings:

$$(1) \qquad Y_i = \sum_t P_{Gt} G_{ti} + \sum_r P_{Nr} N_{ri} = G_i + N_i$$

where Y represents earnings, G_t are the genetic endowments and N_r are the environmental endowments, P_G and P_N are their respective prices. Since G_t and N_r are unobserved, we can express them in units such that each nonzero P is 1. We can also add up the G_t and N_r into G and N which can be thought of as indices.[8]

Below we will discuss the appropriateness of this linear, additive representation, but we will proceed with (1) for the moment. As shown in (2), the variance in Y for individuals equals the sum of the variances of G and N for individuals plus twice the covariance.

$$(2) \qquad \sigma_Y^2 = \sigma_G^2 + \sigma_N^2 + 2\sigma_{GN}$$

Now denote an *MZ* and a *DZ* sibling by an asterisk and a prime, respectively.[9] We can then write the equations for the cross-sib covariance as

$$(3) \qquad \sigma_{YY^*} = \sigma_{GG^*} + \sigma_{NN^*} + 2\sigma_{GN^*}$$

$$(4) \qquad \sigma_{YY'} = \sigma_{GG'} + \sigma_{NN'} + 2\sigma_{GN'}$$

As they stand, equations (3) and (4) have introduced two new observed covariances and six new unknowns. However, it is possible to express many of these new unknowns in terms of the elements in equation (2). Consider first *MZ* twins. Since $G = G^*$, we know that $\sigma_{GG}{}^* = \sigma^2{}_G$ and $\sigma_{GN}{}^* = \sigma_{GN}$. Now let $\rho = \sigma_{NN}{}^*/\sigma^2{}_N$. Then we can rewrite (3) as

$$(3') \qquad \sigma_{YY^*} = \sigma_G^2 + \rho\sigma_N^2 + 2\sigma_{GN}$$

To obtain a comparable equation for *DZ* twins, it is necessary to make a number of strong assumptions. As indicated below, our results are sensitive to several of these assumptions, which unfortunately are not testable within the context of this model. However, the author and Terence Wales have developed a technique which allows us to test these assumptions once we embed the variance component model in a latent variable model.

At this point we will state the assumptions. Later we discuss their validity and implications of relaxing them. First it is assumed that all gene effects are additive, that there is random mating, and no sex linkages. As shown in Oscar Kempthorne, these conditions imply that $\sigma_{GG'} = \frac{1}{2}\sigma^2{}_G$.[10] Second it is assumed that $\sigma_{NN'} = \sigma_{NN}{}^* = \rho\sigma_N{}^2$ or that correlation in the brothers' environments are the same for both types of twins. Since the *DZ* brothers' environments are correlated, we can write $N' = \rho N + z'$ where z' is a random variable. We then assume that G is uncorrelated with z' or that one *DZ* brother's genes are uncorrelated with his sib's specific environment. This assumption implies that

---

[7] We are ignoring nonpecuniary rewards.

[8] The following results on the partitioning of variance are usually derived for the case of a single gene combining at one locus. Kempthorne shows that the same results apply—in the model we use—in the many genes/many loci case provided that there is no epistacy, which means that the effect of each gene is independent of each other gene. See also M. Kendall and A. Stuart, Ch. 36.

[9] In this section I follow Arthur Goldberger, who has greatly simplified and made the author's original presentation of this model more comprehensible.

[10] Suppose we convert each type of gene into a numerical score. Let the father's numerical score, on his two halves of a gene, be F_1 and F_2 and the mother's score be M_1 and M_2. Since an offspring receives one of each pair of his genes from each parent, assuming additive gene effects, his score is $1/2(F_1+F_2)+u$ and $1/2(M_1+M_2)+v$. Calculated over the j families, $\sigma_{GG'}$ is equal to $1/4\sigma_F^2+1/4\sigma_M^2+[1/2\sigma_{MF}+\sigma_{uv'}]$ since F_1+F_2 are independent of v' and M_1+M_2 are independent of u'. The term in brackets is zero if there is random mating. Then assuming $\sigma_F^2=\sigma_M^2=\sigma_G^2$, $\sigma_{GG'}=1/2\sigma_G^2$. Nonadditive genetic effects mean that the average of, say, M_1 and M_2 is not half their sum.

$\sigma_{GN}'=\rho\sigma_{GN}$. With these three strong assumptions we can rewrite (4) as

$$(4')\qquad \sigma_{YY'} = \tfrac{1}{2}\sigma_G^2 + \rho\sigma_N^2 + 2\rho\sigma_{GN}$$

Our model consists of equations (2), (3′), and (4′).

For some purposes, it is more convenient to use correlation coefficients obtained by dividing through by σ^2_Y. In this format our model is

$$(2')\qquad 1 = g^2 + n^2 + 2rng$$

$$(3'')\qquad c^* = g^2 + \rho n^2 + 2rng$$

$$(4'')\qquad c = \tfrac{1}{2}g^2 + \rho n^2 + 2\rho rng$$

where $g^2=\sigma^2_G/\sigma^2_Y$, $n^2=\sigma^2_N/\sigma^2_Y$, $r=\sigma_{GN}/\sigma_G\sigma_N$, $c^*=\sigma_{YY}{}^*/\sigma^2_Y$ and $c=\sigma_{YY'}/\sigma^2_Y$.

This model, which is underidentified, contains three equations and four unknowns—g, n, r, and ρ. The observable parameters are c^* and c. With this model we can estimate $(1-\rho)n^2$ as $1-c^*$; an estimate which incidentally does not use (4″) or the strong assumptions on which it is based. We can interpret $(1-\rho)n^2$ as the proportion of the variance arising from *noncommon* environment of the twins. The *common* environment presumably arises because of family and neighborhood effects, though it will understate the importance of these effects to the extent that identical twins are treated differently by family and friends.[11] Since neighborhoods are generally chosen by parents, we will often call $\rho\sigma^2_N$ the family environment effect.

The difference in the sib correlations is c^*-c, which equals:[12]

$$(5)\qquad c^* - c = \tfrac{1}{2}g^2 + 2(1-\rho)rng$$

Some twin studies impose the added restriction that $r=0$, whence $g^2=2(c^*-c)$. I see no reason to impose that restriction. Instead we will specify values of ρ and solve our system for the other parameters. The analysis is restricted to feasible estimates, defined as $0\leq g, n, r, \rho\leq 1$.[13] These conditions require little comment except those for r. If parents provide environment to compensate for genetic defects (say, eyeglasses or insulin), r will be negative. But we would expect that positive correlation for r would be much more important with more schooling provided to brighter children, for example, or family income which is partly related to genes used to supply the offspring with more schooling and training.

Now let us return to the assumptions that were made to obtain (4′). The first assumption is random mating. Gary Becker recently suggested there may be positive assortive mating, which has been found in empirical studies for education and IQ. However, empirical studies for personality traits such as extroversion have found negative assortive mating.[14] If there is positive assortive mating, then the coefficient on g^2 will be greater than $\frac{1}{2}$ in (4′). Thus an estimate of g^2 based on equation (5) would understate genetic effects.

A second crucial assumption is that ρ is the same for *MZ*s and *DZ*s. Some people argue that ρ_{MZ} is greater than ρ_{DZ}. Indeed, in principle it is possible for such differences to explain fully c^*-c, so that g^2 would be zero. However, in my 1976 paper it is shown that not controlling for genetics in an earnings equation biases downward the coefficient on schooling. While this implies that g^2 cannot be zero, it does not prove ρ_{MZ} equals ρ_{DZ}. Since environment is not defined, it is difficult to assess the argument. There is some evidence in H. Koch,

---

[11] In our sample, common environment may also include some of the effects of military service.

[12] One can also estimate g^2 from $1-c$ or $2c^*-1$. It can be shown, however, that given a certain condition, which holds in our sample, we can do better using c^*-c.

[13] We can write $(1\text{-}2c) = [(1\text{-}2\rho)\ (1\text{-}c^*)/(1\text{-}\rho)] + 2(1\text{-}2\rho)rng$ which can be solved for rng for any ρ except ½. We then solve (5) for g and obtain n from $(1\text{-}\rho)n^2$.

[14] See S. Vandenberg (1972) for some examples for both IQ and personality traits.

for example, that parents tend to dress *MZ* twins more alike on a given day. Of course day to day variation may not indicate much about permanent environment, which is probably more important for the formation of skills and values, etc. In addition, the observed differences in dress or in general treatment of twin sibs may represent parents responding to genetically based needs.[15] If $\rho_{DZ} < \rho_{MZ}$, our estimates for g^2 will be biased upwards. As an aid to the reader in assessing the effects of ρ varying by twin type, we will present some results with $\rho_{DZ} = .9\rho_{MZ}$.

The third important assumption is that $\sigma_{Gj'}$ is zero. While it does not seem patently unreasonable to assume z', one *DZ* brother's specific environment, is uncorrelated with his sib's genes, we have not imposed the same condition on the *MZ*s. If we were to impose this same condition on the *MZ*s and individuals, then in equation (5) $c^* - c$ would equal g^2 and $1 - c^*$ would still equal $(1-\rho)n^2$, but we would not be able to identify the other two parameters. In general this third assumption implies that (5) would be a lower bound to g^2.

The model incorporates several other assumptions that are testable. One of these is that σ_G^2, σ_N^2, and σ_{NG} are the same for both types of twins. The major reason for questioning this assumption is that as a percentage of births, *DZ* twins occur more frequently among older women who generally have larger families, while *MZ* twins are a random event. However, if the two variances and the covariance are the same, then σ_Y^2 should be the same for both twin types. If the *Y*s are normally distributed, we can test the null hypothesis of the same variance with an *F* test.

The model also assumes that *Y* is described by the linear, additive representation in (1) rather than some interactive form such as (6).

$$Y = G + N + c(GN)^d \tag{6}$$

J. L. Jinks and D. W. Fulker show that it is possible to test the null hypothesis that *c* and *d* are zero and that the linear additive form is valid. Their test is to regress for *MZ* twins the absolute value of $Y - Y'$ on a pair's average value of *Y*

$$|Y - Y'| = a + b(Y + Y')/2 \tag{7}$$

They show that if *b* is not statistically different from zero, the linear additive form is appropriate.[16]

III. The *NAS-NRC* Twin Sample

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 Seymour Jablon et al., 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

---

[15] S. Scarr-Salapatek presents some evidence consistent with this view.

[16] Their test is based on the following idea. If equation (1) is valid, $|Y\text{-}Y'|$ equals $|N\text{-}N'|$. But if (6) is valid, $|Y\text{-}Y'|$ equals $|N\text{-}N'| + cG^d(|N\text{-}N'|)^d$. This last term scales the absolute difference in environment by G^d, for which $(Y+Y')/2$ is a proxy. This test assumes that $|N\text{-}N'|$ is distributed homoskedastically.

> 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.[17] [p. 134]

Having received permission from the *NAS-NRC* to contact the twins, we mailed our survey on April 15, 1974 to 12,500 twins 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.[18] In the second mailing on 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 excluded cases where one brother died before being eligible to serve in the military or was exempt from service because of mental and physical defects.[19] 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.[20] All these examples suggest that our sample has 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.[21] R. Cederlöf et al. made a detailed comparison of such questions with

---

[17] I 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 R. Woodworth.

[18] 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.

[19] For an indication of the reasons for and incidence of rejection, which was about 15 percent in World War II, see S. Stouffer et al. Incidentally, there is some evidence in this book that criteria for rejection differed widely by inductee camp and that the distributions are underrepresented in the left-hand tail but not completely omitted.

[20] See K. Christiansen.

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

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, Cederlöf et al. found that the peas in a pod questions agreed with the genetic information about 92 percent 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 subsample of the *NAS-NRC* twins and concluded that the twins' self-assessment of zygosity was correct in 93 percent 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 percent misclassification error will cause the genetic effect to be understated by about 10 percent and a 10 percent misclassification will cause a 20 percent understatement.

From our questionnaire and from data collected by others and entered into the *NAS-NRC* master file, we have obtained information on a wide variety of items, a complete set of which is available on request. The data available include: the twin respondent's 1973 earnings,[22] his hours worked, his wife's earnings, family income, the twin's occupation in 1967 and the socioeconomic status of his initial civilian job, where he was born and a post-World War II residence history, number of children, and a lifetime health history for all different diseases. The sample also contains information about the respondent's family background including parents' education and occupation and number of siblings. Finally we have recently received from Social Security (in a form that protects confidentiality) annual earnings from 1951 to 1974. These data, of course, are only for people in covered occupations and only includes earnings up to the taxable ceiling.

The means and variances of earnings, schooling, and several other variables are given by twin types in Table 1. The average 1973 earnings and education in our sample are \$18,200 and 13.4 years. In the population as a whole, the corresponding figures for white veterans of the same cohort are about \$12,000 and 12 years. About one-quarter of the earnings differential can be eliminated if we reweight our sample observations by parental education and region of birth so as to produce for these variables the average for white males born during the period 1917–27. The sample's average years of schooling is high in part because less than 5 percent of the respondents have less than nine years of schooling.

Next let us compare *MZ*s and *DZ*s. 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 is somewhat greater, although none of the differences are statistically significant. The religious distributions are also very similar for *MZ*s and *DZ*s, which is a bit surprising since I would have expected Catholics 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 percent.[23]

The conclusion that *MZ*s and *DZ*s are random drawings from the same population is further strengthened by a comparison of the simple correlations given in Table 2. In the left-hand portion of that

---

[22] We have another question on earnings that asks pay per normal specified pay period. We have used this to edit the 1973 earnings. Details are available on request.

[23] For samples of this size, the percent level of significance in an F-test is about 1.2.

TABLE 1—SOME SUMMARY STATISTICS FOR INDIVIDUALS IN THE *NAS-NRC* SAMPLE (Calculated separately for *MZ* and *DZ*)

	MZ		*DZ*	
	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.	445
Years of schooling	13.5	9.1	13.3	9.8
Initial full-time civilian 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
Percent Catholic	26	19	23	18
Percent Jewish	4	4	5	5
Percent 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 nonzero for both brothers. For other variables, if one brother answered and the other did not, nonrespondent is set equal to his brother. If both did not answer, both are set at mean or put in "other category." Less than 10 percent of the numbers were estimated. For mother and father data, if brothers' answers differ, mean of responses is used.

[a] Thousands of dollars.

[b] Millions of dollars.

[c] As recalled in 1974.

[d] The Duncan scale.

table, which treats both brothers as individuals, the results are close for *MZ*s and *DZ*s. 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.

TABLE 2—INDIVIDUAL AND CROSS-SIB CORRELATIONS[a]

	Individuals		Cross-Sibs	
	S^{b}	*ln* Y^{c}	S^{b}	*ln* Y^{c}
*MZ*s				
S	1	.44	.76	.40
ln Y		1		.54
*DZ*s				
S	1	.44	.54	.29
ln Y		1		.30

[a] Individual correlations defined as $\sigma_{S\,lnY}/\sigma_S\sigma_{lnY}$. Cross-sib correlations defined as $\sigma_{SY'}/\sigma_S\sigma_Y$, where $\sigma_S lnY'$ is the covariance between one brother's *S* and his sib's *lnY*.

[b] *S* is years of schooling.

[c] *lnY* is *ln* of annual earnings for 1973.

IV. Empirical Results

Our model is based on earnings (Y) being determined by a linear, additive equation. As we indicated, we can test for interactions and functional form by regressing $|\Delta Y|$ on $\overline{Y}$ for *MZ*s. The relevant equations for Y_{73} and $ln\ Y_{73}$ are

$$(7)\quad |\Delta Y| = \underset{(8.2)}{-3252} + \underset{(25.3)}{.545}\overline{Y} \qquad \overline{R}^2 = .39$$

$$(8)\quad |\Delta lnY| = \underset{(1.4)}{.337} + \underset{(.08)}{.0021}(lnY + lnY')/2 \qquad \overline{R}^2 = -.001$$

In the linear equations, the coefficient on $\overline{Y}$ is highly significant with a t-value of about 25. Assuming the distribution of ΔY is homoskedastic, such a finding would lead us to reject the null hypothesis that there is no interactive effect of *G* and *N*. Alternatively we can reject the homoskedasticity hypothesis. In the double *log* form, the coefficient on $(ln\ Y+ln\ Y')/2$ is .002 with a t-statistic of .08 and an R^2 of about .002. Thus at the 5 percent level, we cannot reject the hypothesis of homoskedasticity nor the hypothesis that *ln* Y is a linear, additive function of G and N. While we will rely mostly on the *ln* formulation, we will also present the results for earnings. A similar analysis for years of schooling yielded a positive and significant coefficient in the linear equation, and a negative and significant coefficient in the double *log* equation.

Table 3 contains the main results for 1973 earnings, the *ln* of 1973 earnings, and years of schooling. Since the results for earnings and its *ln* are quite close and since the *ln* form has been shown to be preferable, we will concentrate on its results.

As we noted earlier, the estimate of the impact of noncommon environment $(1-\rho)n^2$ is given by $1-c^*$ where c^* is the cross-sib correlation given in Table 2. For the *ln* of earnings around age 50, we estimate the (additive) contribution of noncommon environment as 46 percent. To divide the remaining 54 percent into genetic, family environment, and covariance terms, we solve for all specific values for ρ for which $0 \leq g, n, r, \rho \leq 1$. The left-hand part of the table is our standard model in which $\rho_{DZ} = \rho_{MZ}$. Looking at the results for the standard model first, we see that environment may account from 50 to 63 percent of the total variance in *ln* Y. Unfortunately, the estimated ranges for the other parameters are not so narrow. Family environment (ρn^2) ranges from 4 to 18 percent, genetics (g^2) ranges 6 to 50 percent, and $2rgn$ varies from 0 to 30 percent.

As shown in the table, the r^2 between genetics and environment ranges from 0 to almost .6. Arbitrary restriction of r^2 to .003 to .11 will limit substantially the ranges of the other parameters. These "more plausible" estimated ranges are 18 to 41 percent and 8 to 15 percent for genetics and family environment, respectively.

It is not clear how to allocate the covariance term between genetics, family, and other environments. But practically any such allocation would indicate that

TABLE 3—CONTRIBUTIONS OF GENETICS, FAMILY, AND OTHER ENVIRONMENTS TO Y, *ln* Y, AND S

Calculated	Standard Model						If $\rho_{DZ}=.9\rho_{MZ}$		
for $\rho=$	g^2	$(1-\rho)n^2$	ρn^2	n^2	$2rng$	r^2	ρ_{MZ}	g^2	n^2
For *ln* Y_{73}									
$\rho=.086$	.50	.46	.04	.50	0	0	.11	.49	.51
.09	.50	.46	.05	.50	.003	.00001	.15	.44	.53
.15	.41	.46	.08	.54	.05	.003	.20	.35	.57
.20	.31	.46	.11	.57	.12	.02	.25	.25	.61
.25	.18	.46	.15	.61	.22	.11	.31	.07	.66
.28	.06	.46	.18	.63	.30	.58			
For Y_{73}									
$\rho=.09$	.50	.45	.04	.49	.005	.00001	.11	.49	.51
.10	.49	.45	.05	.50	.01	.00001	.15	.44	.53
.15	.42	.45	.08	.53	.05	.003	.20	.36	.56
.20	.32	.45	.11	.56	.12	.019	.25	.25	.60
.25	.19	.45	.15	.60	.21	.10	.32	.04	.66
.28	.07	.45	.18	.63	.30	.48			
For Years of Schooling									
$\rho=.5701$	.46	.24	.30	.54	0	0	.62	.34	.61
.57	.45	.24	.30	.53	.01	.0002	.61	.21	.59
.56	.36	.24	.28	.52	.10	.02			
.55	.23	.24	.27	.51	.25	.14			

Note: Entries in body of table are solutions for equations (2′), (3′), and (4″) given the values of ρ in the left-hand column.

genetics plus family environment account for 30 to 55 percent of the total variance, with the estimates of their separate contributions moving in opposite directions as ρ varies. Nearly identical results are obtained if family income or family income minus wife's earnings are substituted for respondent's earnings. From my own viewpoint of equity, I find these estimates large and disturbing.

To provide a basis of comparison, years of schooling has been subjected to the same analysis.[24] We find (the additive part of) noncommon environment accounts for 24 percent of the total. Once again the ranges for n^2 of 51 to 54 percent are limited, as are those for ρn^2 of 27 to 30 percent. But g^2 of 23 to 46 percent and 2 rng of 0 to 25 percent are fairly wide ranges. Still, it appears that family environmental effects are more important for schooling than for earnings.

That the variance of schooling is more determined by common or family environment than that of earnings seems quite reasonable since parents help in choosing the level of education of each brother and provide the same potential resource base. Earnings at age 50, on the other hand, will be influenced by adult events such as on-the-job training opportunities, the *ex post* changes in supply and demand for particular occupations, health, etc., which can differ for the brothers as adults and which influence earnings but not schooling.

There are several assumptions to which our results may be sensitive. One of these is $\rho_{MZ}=\rho_{DZ}$. The right-hand part of Table 3 reanalyzes the model when $\rho_{MZ}=.9\rho_{DZ}$. As expected in this new version, genetic effects are smaller. For earnings, the numerical differences are not that large. For schooling, in which ρ is much bigger, this respecification has bigger effects.

The ranges given in Table 3 are based on a given set of cross-sib correlation coefficients, which are, of course, only sample estimates. If $ln\ Y$ is normally distributed, the standard error of the cross-sib correlation is given by $[(1-R^2)/T]^{1/2}$ where R^2 is the cross-sib coefficient of determination and T is the degrees of freedom. For $ln\ Y$ the standard error is .023 and .027 for MZs and DZs, respectively. If c^* were one standard deviation greater and c one standard deviation smaller, our estimate of $(1-\rho)n^2$ would be decreased by .02 to .44. The maximum value of g^2 would be raised to .57. The other estimates of g^2 would be increased by .10 or less.

V. Conclusions and Caveats

In this paper we have indicated how twin data can be used to partition the variance of earnings (or any other variable) into its genetic, common environment and specific environment proportions. We have also indicated the possible biases and ambiguities in these proportions and have indicated how nonnegativity restrictions can be used to restrict the estimates of underidentified parameters to ranges which are sometimes narrow, sometimes wide.

The paper introduces the *NAS-NRC* twin sample, which I anticipate will be one of the major sources of data on earnings and other variables. Using this data and the methodology presented in the paper, we have obtained the following types of results. We can estimate the contribution of noncommon environment to be 46 percent for the *ln* of earnings and 24 percent for the years of schooling. Making a number of assumptions, we can partition the remaining variance. Using our most plausible estimates, the partitioning of the variance of the *ln* of earnings suggests 18 to 41 percent was due to genetics and 8 to 15 percent to common environment. (The covariance between genetics and environment accounts for the remainder.) For years of schooling the corresponding fig-

---

[24] Neither the genetic nor the environmental indices need be the same for S or lnY.

ures are about 40 percent and 30 percent. However, the feasible ranges of these estimates are much greater.

The next question, of course, is whether the effects of genetics and common environments are large or small. This is a difficult question to answer for the following reasons. First, the variance in environment can change. Since we cannot determine whether or not the environmental variation in this cohort was typical, we cannot determine if our results occur because of an atypical variance in environment. Second, there are no other studies in economics with which direct comparisons can be made. Twins, of course, have been used to study various illnesses and IQ. Arthur Jensen finds g^2 to be about 75 percent for IQ. In this sample, William Pollin et al. find their equivalent to g^2 ranges from 0 for multiple sclerosis to almost 20 percent for schizophrenia. These estimates are from models in which r is assumed to be zero. When we make this assumption, g^2 is 50 percent for earnings and 46 percent for education. This suggests that for earnings and education, genetics has a smaller impact than on IQ but a larger impact than on individual diseases.

Perhaps the judgement should be made another way. *Given* the environmental variation that existed in the cohort under study, a substantial portion of the variance in earnings was determined prior to entry into the labor market. This proportion depends on the value of r^2 selected but will be 8 percent to 41 percent from σ_G^2, plus some fraction of the 5 to 22 percent contribution of $2rng$, plus the 5 to 15 percent from ρn^2. It is possible that the proportion of the variance due to events transpiring before entry into the labor force accounts for half the total variance. While the variances in common and specific environment are not immutable, it still seems to me that for this cohort a large portion of the total variance was determined very early and the effects of individual effort appear to be more limited than I would like to believe.

A. *Caveats*

Given the types of problems we are examining, it is very important to keep in mind the caveats that apply to the twin method in general and this sample in particular. We were able to estimate population variances and covariances because we imposed certain restrictions, such as no assortive mating, no sex-linked genes, and no dominant and recessive genes. Other restrictions which may be suspect are that ρ is the same for DZ and MZ pairs and one DZ's sib's genes are not correlated with his brother's specific environment. In future work we plan to test and, if necessary, relax these assumptions.

Subject to these qualifications, the twin data yield estimates that are applicable to the population of which the twins are representative, which is white males born between 1917 and 1927 who served in the military. In the World War II men were rejected for service because of mental and physical defects which arose from a variety of genetic and environmental causes. It is not clear if the proportions of the variance due to genetics, etc., would be higher or lower than for all white males about 50 years old, because it is not obvious if the truncation was more severe with respect to genetic endowments or environment. A further complication is that the better educated and more successful earners were more likely to respond to our questionnaire.[25] It is not obvious, however, whether the MZ's or DZ's cross-sib correlations have been affected more by response bias.

---

[25] Average earnings in our sample exceed 1971 earnings of veterans by \$7000, which is more than can be accounted for by inflation. The education level of our twins also exceeds that of veterans.

Thus we cannot say if this bias has increased or decreased g^2.[26]

There are several other caveats, which we have not yet discussed, that apply primarily to the generalizability of our results for all white males. To clarify these issues it is best to rewrite equation (1) with prices not standardized to equal 1 and with time subscripts. That is,

$$Y_t = P_{Gt}G_t + P_{Nt}N_t \tag{1'}$$

In order for our results to apply to white males in the United States, at all times it is necessary that $P_{Gt}=P_G$ for all t, for $P_{Nt}=P_N$ for all t, for $\sigma^2_{Gt}=\sigma^2_G$ for all t, for $\sigma^2_{Nt}=\sigma^2_N$ for all t, and $\sigma_{GtNt}=\sigma_{GN}$ for all t. Certainly prices can change if supply or demand for any relevant skills shifts. If P_{Gt} and P_{Nt} do not change proportionately, the contributions of G and N to σ^2_Y will alter. Even with prices fixed the distribution of environment can change; for example, the distributions of schooling, of family size, and size of city of upbringing have altered during this century. Similarly, the distribution of genetic endowments can still be changing because of "recent" successful mutations, changes in environment that alter the advantage of particular gene combinations, or migration. These three reasons suggest that it would be desirable to study twins from other cohorts.

It is also necessary to examine this and other samples at different ages since skills can change because of on-the-job training or because of mental and physical deterioration. We plan to examine earnings of this group at earlier times using Social Security data.

There is one other major caveat. Recently, Jensen and others have been using twin and other studies to determine the heritability of IQ, the extent to which education or environment can raise IQ, and the causes of the difference in average IQ between blacks and whites.[27] Our results have absolutely no bearing on the question of the fitness of the races for earning a living because differences in average earnings between blacks and whites can be due to discrimination as well as the average differences in genetic endowments and environment. Moreover, even if one were to adopt the view that there is no discrimination in the labor market,[28] we cannot partition average racial differences in earnings into differences arising from environment and from genetic endowments until we know the average difference in G and N. But with few exceptions, both the G and N that are relevant to earning a living are still mostly unmeasured and undefined.

I began this paper by arguing that it was important to estimate the effects of genetics, family, and other environments because of equity considerations. Assuming the results in this paper are at all close to the mark for this and more recent cohorts, it would seem that much of the inter- and intragenerational inequality of earnings is related to who one's parents are. Moreover, I suspect that these family effects will not be overcome by equality of opportunity programs designed solely to eliminate market imperfections. Transfer and

---

[26] For occupational status we have responses for some pairs who did not respond to our survey. The results for nonresponders and responders for this variable are in rough accord though the estimate of g^2 is greater for responders to our survey. However, in some preliminary work in which we reweight the sample to obtain population weights for parental education and region of birth, our estimates of cross-sib correlations are unchanged, though our estimates of mean earnings and education are lower.

[27] Jensen has written me the following about this paragraph: " . . . It is not quite correct to claim I use twin studies to determine the causes of the average differences between the races. I think there is some relevant connection, but it is far from direct, as your statement would imply" (Letter dated Nov. 6, 1974).

[28] See Becker (1971) and Arrow for reasons why discrimination might vanish in the long run.

other programs can be used to achieve greater equality of outcome whether the source of the inequality is genetic, family, or other environment.

REFERENCES

K. Arrow, "The Theory of Discrimination," in O. Ashenfelter and A. Rees, eds., *Discrimination in Labor Markets,* Princeton 1973.

G. Becker, "A Theory of Marriage, Part I," *J. Polit. Econ.,* July-Aug. 1973, *81,* 813–46.

———, *The Economics of Discrimination,* 2d ed., Chicago 1971.

L. Cavalli-Sforza and W. Bodmer, *The Genetics of Human Populations,* San Francisco 1971.

R. Cederlöf et al., "Studies in Similarity Diagnosis in Twins with the Aid of Mailed Questionnaires," *Acta Genet. Statist. Med. Basel,* 1961, *11,* 338–62.

K. Christiansen, "Mobility and Crime Among Twins," *Int. J. Criminology and Penealogy,* Feb. 1973, *1,* 31–45.

D. Falconer, *Introduction to Quantitative Genetics,* New York 1960.

R. A. Fisher, "The Correlation between Relatives on the Supposition of Mendelian Inheritance," *Transactions of the Royal Society, 52,* Edinburgh 1918, 399–433.

A. Goldberger, "On Taubman's Model for Twins," mimeo, Univ. of Wisconsin 1975.

→ **A. Hoffer and W. Pollin,** "Schizophrenia in the NAS-NRC Panel of 15909 Veteran Twin Pairs," *Arch. Genetic Psychiatrics,* Nov. 1970, *23,* 469–78.

S. Jablon et al., "The NAS-NRC Twin Panel: Methods of Construction of the Panel, Zygosity Diagnosis and Proposed Use," *Amer. J. Hum. Genetics,* Mar. 1967, *19,* 133–61.

→ **A. Jensen,** "Estimation of the Limits of Heritability of Traits by Comparison of Monozygotic and Dizygotic Twins," *Proc. Nat. Acad. Sci.,* 1967, *58,* p. 149.

→ **J. Jinks and D. Fulker,** "Comparison of the Biometrical Genetical, MAVA, and Classical Approaches to the Analysis of Human Behavior," *Psychol. Bull.,* May 1970, *73,* 311–49.

L. Kamin, "Heredity, Intelligence, Politics and Psychology," mimeo, Princeton Univ.

O. Kempthorne, *An Introduction to Genetic Statistics,* New York 1957.

M. Kendall and A. Stuart, *The Advanced Theory of Statistics,* Vol. 3, London 1966.

H. Koch, *Twins and Twin Relations,* Chicago 1966.

P. Mittler, *The Study of Twins,* Middlesex 1971.

W. Pollin et al., "Cycle Pathology in 15,909 Pairs of Veteran Twins," *Amer. J. Pysch.,* Nov. 1969, *126,* 597–610.

S. Scarr-Salapatek, "Twin Method: Defense of a Critical Assumption," mimeo, Univ. Minnesota 1975.

H. Scheffe, *The Analysis of Variance,* New York 1959.

→ **T. Schultz,** *Investment in Education: The Equity-Efficiency Quandary, J. Polit. Econ.* suppl., June 1972, *80,* S1–S292.

S. Stouffer et al., *The American Soldier,* Vol. 4, Princeton 1950.

P. Taubman, *Sources of Inequality of Earnings,* Amsterdam 1975.

——— "Earnings, Education, Genetics and Environment," *J. Hum. Resources,* Fall 1976.

——— **and T. Wales,** *Higher Education: An Investment and a Screening Device,* New York 1974.

——— **and** ———, *Mental Ability and Higher Educational Attainment in the 20th Century,* Occas. Paper 118, Nat. Bur. Econ. Res., Berkeley 1972.

→ **S. Vandenberg,** "Assortive Mating or Who Marries Whom," *Behavior Genetics,* June 1972, *2,* 127–57.

——— "In Defense of Twins and the Twin Method," in *Essays in Honor of Professor Essen Moller, Acta Psychiatrica Scandinavica,* 1970 suppl., *46,* 219.

R. Woodworth, *Heredity and Environment,* Soc. Sci. Res. Council Bull., *47,* New York 1941.