American Educational Research Journal
Summer 1980, Vol. 17, No. 2, Pp. 153–170

Uses of Sibling Data in Educational and Psychological Research

ARTHUR R. JENSEN
University of California, Berkeley

Methods are explained, with empirical examples, for using sibling data on psychometric variables (1) as a covariate for statistically controlling family background in psychological and educational experiments, (2) as a means for testing the adequacy of age-standardized scores, (3) for testing the interval scale property of mental measurements, (4) for analyzing correlations into between- and within-family components and distinguishing intrinsic from extrinsic correlations between variables, and (5) for detecting cultural (i.e., between families) sources of variance in psychological tests.

Researchers generally treat data obtained on related persons, family members, or various kinships in general, as pertinent only to research in genetics or behavioral genetics. Kinship data are, of course, the main grist of research in human genetics (e.g., Erlenmeyer-Kimling & Jarvik, 1963). But kinship data can also serve useful purposes in psychometrics and in psychological and educational research, quite apart from any concern with genetical analysis per se (e.g., Jensen, 1974; 1977). The aim of this paper is to explain five such uses of sibling data, with examples based on data obtained by the author.

The most plentiful and easiest kinship data to obtain in the school-age population are data on full siblings. A substantial proportion (we have found proportions from about .50 to .70 in various California schools) of school children have one or more siblings enrolled in the same school. Of course, the proportion of children with siblings in the same school is larger for

---

This article is based on a paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

children in the middle grades, so that is where one should begin to recruit subjects for a sibling study.

(1) SIBLING DATA AS A COVARIATE CONTROL

In educational experiments we frequently wish statistically to control independent variables in order to improve the statistical detection of treatment effects. Pupils' intelligence or prior level of achievement are commonly used covariates for this purpose, as are measures of family background. Sibling data may be used as an additional measure of family background factors as they affect the particular dependent variable in the experiment, for example, scores on a test of scholastic achievement. The covariance (or correlation) between the age-standardized achievement scores of the pupils in the study proper and of their siblings (preferably the siblings closest in age) represents variance due to "family background" factors common to siblings from the same family. Or, to put it another way, it is variance due to differences *between* families. Using nearest-in-age siblings, measured on the same age-standardized dependent variable, as a covariate control, along with other subject variables, such as IQ or the subject's socioeconomic status, which may also be included as covariates in the ANCOVA, may increase the power of the experiment. This is especially true in the abilities domain, where the correlations between siblings are substantial. (For example, the mean of the sibling correlations for IQ reported in the literature is +0.49.) Sibling data on the dependent variable is often easier to obtain than any other family background data. This can be an important consideration when there are restrictions on the invasion of privacy, which may be involved in obtaining family background data. Sibling scores on the dependent variable may serve the same purpose, controlling for family background, without the invasion of privacy implied in obtaining other family background information such as parental income, occupation, and education.

"Family background" is an ambiguous and poorly understood term in educational research. It is often mistakenly understood to mean an exclusively environmental source of variance, but, in fact, it is nearly always some composite of *between-families* environmental differences and genetic differences.

Between- and Within-Family Variance. Variance *between* families (ignoring measurement error) is due to those genetic and environmental influences that are common to all siblings *within* a family, but that differ *between* families. ("Family" in this context refers only to full siblings who are reared together.) Variance *within* families (ignoring measurement error) is due to those genetic and environmental influences that cause siblings (reared together) to differ from one another.

Some classes of variables can have both *between* family and *within* family differences. This is true of all genetic factors that cause variation in the

population, except in the case where the siblings are from a monozygotic multiple birth, such as identical twins or triplets, and so forth, for which there is no within-family genetic variance. Some types of variables, such as racial origin, have *between*-family but no *within*-family differences. Still other types of variables have no within-family variation at one period of life but have it at another period; for example, socioeconomic status (SES). Siblings reared together during childhood are considered to have the same SES. But later, as adults, siblings may differ in SES, in terms of their differing amounts of education, occupational levels, and earnings. There are no characteristics that differ *within* families but not *between* families. Any genetically conditional variable on which siblings differ will inevitably show differences *between* families in the next generation. This can be stated in general terms as the *First Law of Differential Psychology*: *All* within-*family phenotypic and genotypic differences and correlations also exist as* between-*family phenotypic and genotypic differences and correlations, but the reverse is neither necessarily nor always true.*

Sibling Correlation. The correlation between siblings on a given variable is entirely attributable to variance *between* families. The sibling correlation is, in fact, the *proportion* of the total variance which is attributable to variance between families. In the analysis of variance, the total variance in the population is partitioned into a *between*-families component σ_B^2 and a *within*-families component σ_W^2. The *population* intraclass correlation between siblings, then, is $\rho_i = \sigma_B^2/(\sigma_B^2 + \sigma_W^2)$. The *sample* intraclass correlation between siblings, derived from the between and within mean squares (MS_B and MS_W) of the analysis of variance, is $r_i = (MS_B - MS_W)/[MS_B + (\bar{k} - 1)MS_W]$, where $\bar{k}$ is mean of the frequency distribution of the number of siblings per family (Haggard, 1958):

$$\bar{k} = \frac{N - \sum k^2/N}{F - 1},$$

where N = total number of subjects,
k = number of siblings in each family, and
F = number of families.

The sibling *intraclass* correlation r_i estimates the correlation between siblings in general, since, unlike the *interclass* or Pearsonian r, it does not assign siblings to different classes such as x and y. The simple Pearsonian r_{xy} which would be obtained by correlating, say, younger siblings (as x) with their next older siblings (as y), or lower scoring siblings with their higher scoring siblings, or any other classifications of siblings as x and y, will rarely be the same as the intraclass r_i, which closely approximates the average value of all possible Pearson correlations between siblings.

In genetical research, where the interest is in estimating the proportion of the total variance that a given kinship shares in common, the intraclass

correlation is the proper measure of relationship. In certain psychometric or statistical uses of sibling (or other kinship) data, however, the Pearson *r* is appropriate. This is evident in any use of sibling data that implies a distinct classification of the members of each sibling pair, such as younger versus older, male versus female, higher versus lower scoring, and so forth. Pearson *r* is obviously called for in the use of siblings as a covariate control in the analysis of covariance discussed above, since one member of each sibling pair is classed as the covariate in the ANCOVA. (Pearson *r* is, of course, implicit in the usual computational routines for ANCOVA.)

(2) TESTING THE AGE-STANDARDIZATION OF MEASUREMENTS

Since siblings (except twins) naturally differ in age, all of the uses of sibling data described herein require age-standardized measurements. Most published standardized tests provide age-standardized scores, which can be rigorously tested for adequacy of the age standardization by the use of sibling data, provided the sibling sample is sufficiently large (say total $N > 200$) and representative of the population on which the test was standardized. If the sibling sample is not representative, this fact will be clearly revealed by the analyses described later. If the age standardization is shown by the sibling method to be inadequate and one wishes to use the sibling data for one of the other purposes described in this article, the measurements should be re-standardized, if possible, to remove any artifacts due to age differences between siblings. Poorly age-standardized measurements have the effect of artifactually *inflating* the correlation between twins (who are always the same age) and artifactually *attenuating* the correlation between siblings (who are always of different ages).

There are two main methods for obtaining age-standardized measurements: (a) normalized standardization, and (b) non-normalized standardization. In either method siblings are not required, but neither are they necessarily excluded.

(a) *Normalized Standardization*

This method is advisable only when the total *N* is quite large. The total age range in the sample is divided into as many equal intervals as possible, with the limitation that no interval contain fewer than 100 participants. The equal age intervals should not be greater than 1 year and need not be less than 3 months—at this limit there is more advantage in having larger *N*'s within each age interval than in having narrower age intervals. The raw scores within each age interval are rank ordered from highest to lowest, and the ranks are then converted to percentile ranks. Using the tabled areas under the normal curve, the percentile ranks are transformed to *z* scores, which are normalized standardized scores. (The *z* scale may, of course, be

subjected to any linear transformation to give the scores any mean and standard deviation one deems most convenient for one's purpose.) I have generally found that for ability and achievement tests, normalized standardized scores have the most desirable scale properties; for example, they usually satisfy independent criteria for determining an equal interval scale. (See section 3.)

(b) *Standardized Age-regressed Scores*

With this method one determines the best-fitting regression line (linear or nonlinear) of raw test scores on age (in months), using standard methods of curve fitting. A trend analysis should be done to determine if the regression of raw scores on age is significantly nonlinear.

If the regression of raw scores on age does not depart significantly from linearity, as often happens for ability tests in the age range from about 5 to 13, one obtains the age-regressed scores $\hat{Y}$ for each participant from the simple regression equation. That is, $\hat{Y} = b_{XA}(A - \bar{A}) + \bar{X}$, where b_{XA} is the coefficient of regression of raw scores on age (in months), $\bar{A}$ is the mean age of the entire sample, and $\bar{X}$ is the mean test score in the entire sample. (The values of $\hat{Y}$ can be transformed to z scores or any other convenient transformation; or they can be converted to percentile ranks and then transformed via the tables of the normal curve, to normalized z scores or any desired linear transformation of the normalized z scores.)

If the regression of raw scores on age departs significantly from linearity, as revealed by the trend analysis described below, the age-regressed score $\hat{Y}$ for each individual can be obtained from the regression equation employing the set of statistically significant regression coefficients yielded by the trend analysis.

The trend analysis referred to above is most easily performed by means of a stepwise multiple regression analysis, in which successive powers of age (in months) are entered as the independent (predictor) variables and raw test score is the dependent variable. Powers of age (i.e., age^1, age^2, age^3, etc.) are entered stepwise into the mutiple regression analysis until the increment in R^2 is nonsignificant (at any desired level of confidence α), as determined by the usual F test.

The adequacy of the age standardization of test scores can be most rigorously tested by the use of siblings. Ordinarily one can test the adequacy of the age standardization by testing the significance of the correlation r (or the multiple R, using powers of age as the predictor variables) between age and the standardized scores. The correlation should not differ significantly from zero if the age standardization is adequate. Standardized scores obtained on a large sample of siblings varying in age over the age range of the original standardization sample provide *two* independent tests of the adequacy of the age standardization. The first statistical test is the Pearson

correlation (or multiple R, using powers of age) between the *mean* age of sibling pairs and the test score *means* of sibling pairs (or sets of any number of siblings). This is referred to as the *between*-families correlation r_B between age and test scores. The second statistical test is the Pearson correlation (or the multiple R, using powers of age) between (a) the age difference between older (O) and younger (Y) siblings (i.e., age of O minus age of Y), and (b) the test score difference between older and younger siblings (i.e., score of O minus score of Y). This is refered to as the *within*-families correlation r_W between age and test scores. The expected value of these correlations is zero, under the hypothesis that the test scores have been adequately age-standardized. That is to say, properly age-standardized scores should have zero correlation with age. If either r_B or r_W, or both, differ significantly from zero, the hypothesis of adequate age standardization can be rejected. The combined probabilities of r_B and r_W provide a more powerful test of the hypothesis when r_B and r_W separately have p values greater than α (α is the size of the critical region used, or the probability of making a Type I error, that is, rejecting the null hypothesis when it is true). If we know that exact p values (call them p_B and p_W) or r_B and r_W, we can test the significance of the combined result using the method suggested by Fisher (1970, pp. 99–101); the value $-2(\log_e p_B + \log_e p_W)$ is distributed as chi-square for 4 degrees of freedom.

Correction of BF and WF Correlations for Attenuation

A more stringent test of the hypothesis that the scores are adequately standardized for age is achieved if r_B and r_W can be corrected for attenuation. The reliability of age is assumed to be perfect, and the reliability r_{XX} of the test scores is estimated in the usual way. The reliability r_X of the sibling pair *means*, then, is

$$r_X = (r_{XX} + r_{YO})/(1 + r_{YO}), \tag{1}$$

where r_{YO} is the Pearson r between younger and older siblings. The reliability r_{O-Y} of the *differences* between older and younger siblings is

$$r_{O-Y} = (r_{XX} - r_{YO})/(1 - r_{YO}). \tag{2}$$

Correction for attenuation of r_B and r_W, then, is $r_B' = r_B/r_{\bar{X}}^{1/2}$, and $r_W' = r_W/r_{O-Y}^{1/2}$.

(3) SIBLING TEST OF INTERVAL SCALE

It is often important in a psychological or educational study to have some independent evidence that the measurements constitute an equal-interval scale. For example, in testing the hypothesis that a particular type of

instruction should produce a *larger* gain in achievement scores (as measured against an IQ-matched control group) for low IQ than for high IQ pupils, we cannot meaningfully interpret the lack or presence of this hypothesized *interaction* between IQ level and the magnitude of the experimental effect without some evidence that the dependent variable, achievement, is measured on an equal-interval scale. Usually we simply assume that the trait (in this case, achievement) is normally distributed in the population, construct the test in such a way as to yield a normal distribution of scores (or normalize the scores by some suitable transformation), and then conclude that the scores constitute an interval scale.

Sibling data provide an independent test of the hypothesis that the scores that we wish to interpret as an interval scale are, in fact, an interval scale. The test, in essence, is the correlation r_{MD} between sibling *means* and sibling absolute *differences* on the measurements in question.

In a random sample of the general school population, sibling means vary over a wide range on intelligence and achievement tests. With respect to the hypothesis, a sibling mean indicates the average level of the sibling pair on the measurement scale. We wish to know if this average level of the sibling pairs is significantly correlated with the absolute difference between the siblings. The expected correlation r_{MD} should be zero if the measurements (scores) are an equal-interval scale. This is a compelling inference only if there is no equally compelling theory that siblings should truly differ more in certain parts of the total range of scores than in other parts. Sibling differences are theoretically analyzable into three components of variance: (a) within-family genetic variance, (b) within-family environmental variance, and (c) error variance (i.e., unreliability of measurement). There is nothing in genetic theory that would lead to the expectation of a nonzero correlation between *mean* sibling genotypic values (which reflect between-families genetic variance) on a continuous trait and the *differences* between siblings' genotypic values (which reflect within-families genetic variance), excluding cases of major gene defects (e.g., phenylketonuria) and chromosomal anomalies (e.g., Down's syndrome). Also, there is no *general* theoretical rationale that would lead to the expectation of a nonzero correlation between sibling means and environmental differences among siblings (i.e., within-families environmental variance). Finally, we can empirically determine whether measurement error is homogeneous throughout the full range of the scale of scores.

The test of the hypothesis of an interval scale proceeds as follows:

(*a*) The scale is assumed to be properly age-standardized. The adequacy of the age-standardization can be checked by the method described in section 2.

(*b*) The homogeneity of reliability can be checked, using all of the data without respect to sibling classification. The test is split into equal halves by

some psychometrically sensible method, such as odd versus even-numbered items. The Pearson r (or multiple R) between total scores and the absolute difference between the two half-scores provides a test of the homogeneity of reliability. (This is unfortunately not a powerful test and can often result in a Type II error [i.e., accepting the hypothesis $\rho = 0$ when $\rho \neq 0$], the risk of which can be lessened by setting the significance level for rejecting the null hypothesis at $p < .10$ or even $p < .15$.) The test is performed by the same general method as the test for interval scale described in (*c*) below, except that correction for attenuation is not possible, since one is actually testing the hypothesis that the reliability of the absolute difference scores is zero, that is, that they are purely random errors which cannot be correlated with the true scores.

(*c*) The test for interval scale is most rigorous if one uses stepwise multiple correlation to detect any nonlinear trend, since the inequalities in score intervals may not be a simple linear function of the true scale. To simplify exposition I shall use M for the *mean* of a sibling pair and $|D|$ for the *absolute difference* between siblings. Only two siblings per family should be used, preferably those nearest in age. (Obviously, the test scores should not determine the selection of siblings.) Obtain two stepwise multiple correlations (1 and 2, below) based on the n independent variables which add significant stepwise increments to R^2.

Independent Variables	*Dependent Variable*
(1) $M, M^2, M^3 \ldots M^n$	$\lvert D \rvert$
(2) $\lvert D \rvert, \lvert D \rvert^2, \lvert D \rvert^3 \ldots \lvert D \rvert^n$	M

The resulting multiple correlation coefficient R, in each case, should be corrected for bias, using the well-known "shrinkage" formula:

$$R_c = \sqrt{1 - (1 - R^2)[(N - 1)/(N - n)]}, \tag{3}$$

where N is the number of sibling pairs and n is the number of independent variables. Each R_c should then be corrected for attenuation, to obtain $R_{c'} = R_c/(r_M r_{|D|})^{1/2}$, where r_M and $r_{|D|}$ are the reliabilities of the sibling means and sibling absolute differences, respectively:

$$r_M = (r_{XX} + r_i)/(1 + r_i) \tag{4}$$

and

$$r_{|D|} = (r_{XX} - r_i)/(1 - r_i), \tag{5}$$

where r_{XX} is the test reliability and r_i is the intraclass correlation between siblings. Finally, the two values of R_c are each tested for significance by means of the t test:

$$t = R_c N^{1/2}/(1 - R_{c'}^2), \tag{6}$$

where N is the number of sibling pairs. If t does not fall in the critical region (i.e., $p > \alpha$), we conclude that the test scores are an interval scale over the range subtended by the total distribution of M.

This method was applied in a study (Jensen, 1977) of the California Test of Mental Maturity IQ scores in large sibling samples of white and black school children in rural Georgia who differed by 2σ in mean IQ. The test of interval scale was applied in each racial sample separately, and in the combined samples. In no case could the hypothesis of an equal interval scale be rejected.

(4) BETWEEN- AND WITHIN-FAMILY CORRELATIONS.

Partitioning correlations between variables into their between-families (BF) and within-families (WF) correlations has theoretically important uses in differential psychology and psychometrics. It permits a separation of between-families social-cultural factors (macroenvironments) and within-families (microenvironmental) factors that contribute to the total variance in the population. The BF correlation (corrected for attenuation) reflects genetic and environmental factors that differ *between* families but not among siblings *within* families. The WF correlation (corrected for attenuation) reflects genetic and environmental factors that differ among siblings. In the absence of assortative mating (i.e., correlation between parents), the genetic component of the total variance is evenly divided between BF and WF; that is, half of the total genetic variance in the population exists *between* families and half of it exists *within* families. The increase in genetic variance due to assortative mating all goes into the BF variance, whereas the WF variance is negligibly affected by assortative mating (see Jensen, 1978).

Partitioning Sibling Covariance and Correlation

The total covariance between any two variables X and Y is $\sum xy/N$, where x and y are deviations from the mean (i.e. $x = X - \bar{X}$ and $y = Y - \bar{X}$) and N is the number of paired variables. (Pearson $r = \sum xy/N\sigma_x\sigma_y$.)

If n pairs of siblings, with the members of each pair designated as a and b respectively, are each measured on variables X and Y, the total covariance between X and Y can be analyzed into two additive parts: *between*-families (BF) and *within*-families (WF) covariance. The BF covariance is the mean cross-product of the sibling means on x and y:

$$\text{BF cov} = \frac{1}{n}\sum\left(\frac{x_a + x_b}{2}\right)\left(\frac{y_a + y_b}{2}\right). \tag{7}$$

The WF covariance is the mean cross-products of one-half of the sibling differences (i.e., the mean of the deviations of each sibling from the family mean):

$$\text{WF cov} = \frac{1}{n}\sum\left(\frac{x_a - x_b}{2}\right)\left(\frac{y_a - y_b}{2}\right). \tag{8}$$

The total covariance between X and Y is:

$$\text{Total cov} = \frac{1}{2n}(\sum x_a y_a + \sum x_b y_b). \tag{9}$$

By expanding equations (7) and (8) it can easily be seen that the total covariance is equal to the sum of the BF and WF covariances:

$$\text{BF cov} = \frac{1}{4n}(\sum x_a y_a + \sum x_b y_b + \sum x_b y_a + \sum x_a y_b),$$

$$\text{WF cov} = \frac{1}{4n}(\sum x_a y_a + \sum x_b y_b - \sum x_b y_a - \sum x_a y_b).$$

In working with correlations, it is simplest to use the sum of sibling scores on X and Y to obtain the BF r_{xy}, and the difference between sibling scores to obtain the WF r_{xy}. Obviously the *direction* of the sibling difference must be consistent for X and Y (i.e., $X_a - X_b$ and $Y_a - Y_b$); it is most convenient to assign older and younger siblings to a and b, respectively.

The BF and WF correlations should be corrected for attenuation in the usual way, using the appropriate reliabilities as given in formulas (1) and (2).

Intrinsic and Extrinsic Correlation

A correlation between two variables may be *intrinsic* or *extrinsic*, a distinction that can have considerable theoretical importance. Both intrinsic and extrinsic correlations can have either genetic or environmental components, or both. The distinction between intrinsic and extrinsic correlation is revealed by BF and WF correlations. The relationships between intrinsic and extrinsic correlations and BF and WF correlations are shown in Table I.

TABLE I

Relationships Between Intrinsic and Extrinsic Correlations, Genetic (G) and Environmental (E) Components, and Between-Families (BF) and Within-Families (WF) Correlations

All Possible Combinations of BF and WF Correlations	Type of Correlation	
	Intrinsic	Extrinsic
(1) BF $\rho > 0$, WF $\rho > 0$	G and/or E	
(2) BF $\rho > 0$, WF $\rho = 0$		G and/or E
(3) BF $\rho = 0$, WF $\rho > 0$	Nonexistent Condition	
(4) BF $\rho = 0$, WF $\rho = 0$	Population Correlation is zero.	

All *intrinsic* correlations are characterized in general by two properties: they cannot be wiped out (or reversed in sign) by means of experimental manipulation or by means of genetic selection.

Four main types of correlation can be referred to as *intrinsic*:

(1) *Causal-functional.* Variables X and Y involve a direct causal relationship, such that the experimental manipulation of X is accompanied by a change in Y. *Example:* number of learning trials (X) and amount of retention (Y).

(2) *Common factor.* X and Y are both measures of (or are both correlated with) some common factor. *Example:* strength of left-hand grip (X) and strength of right-hand grip (Y); height (X) and weight (Y); or speed of learning X and speed of learning Y.

(3) *Part-whole.* One variable is some part of the other, or skill X is a subset of skill Y. *Example:* leg length (X) and height (Y); or skill in short division (X) and skill in long division (Y).

(4) *Pleiotropy.* X and Y are both affected by the same gene(s), even though X and Y may be of phenotypically quite distinct characteristics. Variation in both traits is linked to a common (genetic) causal factor. A pleiotropic gene has two (or more) distinct phenotypic effects. *Example:* phenylketonuria, a single-locus genetic defect, results in mental retardation (X) and lightness of skin and hair pigmentation (Y).

Extrinsic correlations are characterized by the fact that, at least in principle, they can be wiped out or even reversed in sign by means of experimental or environmental manipulation or by means of genetic selection.

There are two types of *extrinsic* correlation: genetic and environmental. Extrinsic genetic correlations are of two kinds: (1) nonlinked genetic correlation and (2) correlation by genetic linkage.

(1) *Nonlinked genetic correlation.* Variables X and Y an be correlated in the population through common assortment of genes due to cross-assortative mating for certain characteristics which have no functional or other intrinsic

relationship to one another. Say, for example, there is zero correlation between curly hair and height in the population, but in the next generation there is a strong tendency for tall men to marry curly-haired women (i.e., cross-assortative mating for the two characteristics); then, in the next generation there will be a genetic correlation between height and curly hair in the population, i.e., persons with curly hair will be somewhat taller, on the average, than persons with straight hair. Such cross-assortative mating can create a correlation in the population between any genetically conditioned traits, even though there may be no intrinsic relationship between the traits.

Such genetic correlations (unless there is genetic linkage, which is highly improbable for continuous or polygenic traits) have the important property that they are entirely *between*-families (BF) correlations and have no *within*-families (WF) component. Even though there is reliable WF *variance* on each trait, the expected value of the WF *correlation* between the traits is zero. Hence, a test of the hypothesis that WF $\rho_{xy} = 0$, when the BF $\rho > 0$, decisively rules out genetic linkage between traits X and Y as well as any form of intrinsic correlation between them. If WF $\rho_{xy} = 0$, then the total population $\rho_{xy} > 0$ (i.e., for all individuals in the population) represents entirely *extrinsic* correlation, due either to (*a*) genetic correlation (resulting from cross-assortative mating for the correlated traits), or to (*b*) the extrinsic component of what I refer to as *environmental correlation* (see below), or to a combination of (*a*) and (*b*).

The fact that a correlation is said to be *extrinsic* according to this criterion does not make it any less real or reduce its predictive validity for individuals in the general populations. The distinction between *intrinsic* and *extrinsic* correlation, however, does imply different theoretical interpretations of the correlations so classified. As I have explained elsewhere (Jensen, 1979), experimental or so-called process analysis of correlated tests, or of test scores and certain physical measurements, can lead to a greater understanding of the correlated variables only if there is an *intrinsic* correlation between them. For example, the correlation between height and IQ appears to be entirely extrinsic. I and others (e.g., Laycock & Taylor, 1964) have found a significant positive BF correlation (ranging from about +0.1 to +0.3) between height and IQ, whereas the WF correlation between height and IQ is zero. The fact that the correlation between height and IQ is an *extrinsic* correlation, as indicated by these findings, implies that study of the nature of variation in height can afford no scientific clues to the nature of individual variation in IQ. The discovery of physical correlates of mental traits can possibly lead to a greater understanding of the mental traits, provided the correlation is intrinsic, that is, a WF correlation that is not due merely to genetic linkage. We know, for example, that there is a correlation of about +0.30 between IQ and brain size (VanValen, 1974), but we do not know if this is an intrinsic correlation. It would seem important to find out. If it is only an extrinsic correlation we can forget it, as far as its role in a theory of intelligence is

concerned. The same thing can probably be said about the correlations between IQ and a number of different blood types (Osborne & Suddick, 1971). The positive correlation between lightness of skin pigmentation and IQ in the American black population (studies reviewed by Jensen, 1973, pp. 222–224) may or may not be an intrinsic correlation; no one has yet determined the WF and BF correlations between IQ and skin color. If the WF correlation is zero, it would rule out the hypothesis which explains the observed correlation in the black population in terms of adverse effect on IQ of social prejudice against darker skin. The counter hypothesis would be that the correlation is entirely BF due to cross-assortative mating for skin color and IQ within the black population—a finding that might be of sociological interest but not of any importance in terms of psychological or genetical theory.

Another example is the negative correlation between IQ and delinquency. This turns out to be an intrinsic or WF correlation. Delinquent and nondelinquent siblings of the same sex differ almost as much in IQ as do unrelated delinquents and nondelinquents in the general population (Hirschi & Hindelang, 1977). This implies that IQ somehow mediates delinquency—a quite different (and much more important) theoretical implication than would be the case if only a BF correlation (but not a WF correlation) were found between delinquency and IQ.

(2) *Correlation by genetic linkage.* Variables X and Y may be correlated WF only because they are genetically linked; that is, the genes influencing X and Y are located on the same chromosomes in close proximity to one another, so that they have a higher probability of remaining together in the crossing-over process in gametogenesis (the meiotic formation of the sex cells). The two traits, however, may not be any more intrinsically correlated than in the case of nonlinked genetic correlation; and as the linkages break up in successive generations, the WF correlation (and the BF correlation, too) will gradually decrease toward zero. Linkage would account for little, if any, of the correlation between continuous or polygenic traits, but would have to be considered, for example, in considering correlations between particular single-gene blood groups and IQ. The finding of such a genetic linkage would not be psychologically informative about the nature of IQ, but it would be of considerable importance in the study of the genetic inheritance of intelligence. There are methods in quantitative genetics for the statistical detection of linkage (e.g., Cavalli-Sforza & Bodmer, 1971, pp. 870–880).

Environmental Correlation: Intrinsic and Extrinsic

This is a class of correlations which may be either intrinsic or extrinsic; these two aspects show up in the WF and BF components, respectively, of the overall correlation in the population. If two classes of events, X and Y, occur together more than chance in the individual's environment, through

common interest or exposure, then knowledge of *X* probabilistically (but not causally) implies knowledge of *Y*. *Example:* A person who knows a lot about, say, baseball is also likely to possess above-average knowledge of football, through common interest in sports. An opera lover is likely to have more than average knowledge of symphonies.

Such environmental correlations can exist WF as well as BF. Sex differences in tests of various kinds of information are an example. There is a WF correlation between knowledge of sewing and cooking, and between knowledge of sports and auto mechanics, largely because of the different experiences of males and females. One would expect these WF correlations to be higher in families with opposite-sex siblings than in families with same-sex siblings.

Certain kinds of environmental experiences may be much more highly associated for *all* siblings of some families than for *all* siblings of other families, because of differing family interests, life styles, cultures, and so forth. Measurements of the knowledge or skills derived from such associated experiences will then show a much higher BF correlation than a WF correlation. Marked differences between BF and WF correlations, and particularly a difference between the *pattern* of intercorrelations or the *factor structure* of the BF and WF intercorrelations of a number of variables, indicates that at least some of the intercorrelations among the variables are either extrinsic environmental correlations, or extrinsic genetic correlations, or both. Different social classes or racial groups, for example, might have different commonly associated cultural experiences, which generally affect all members of a family as well as most families in the particular group. The WF and BF correlations between measurements that reflect these experiences will therefore be different in different subpopulations. If variance in mental test scores were largely the result of differences in social class, cultural background, economic privilege, parental education, family values, and the like, as is often conjectured, then we should expect most of the significant intercorrelations among such tests to be BF rather than WF, and it would seem reasonable to expect different patterns (or factor structures) of BF and WF intercorrelations among various tests, and in different subpopulations.

(5) SIBLING TEST OF CULTURE BIAS

The hypothesis that a test is not culturally biased for two or more subpopulations can be rejected if it can be demonstrated, at some acceptable level of statistical significance, that any essential psychometric characteristic of the test differs between subpopulations. The one characteristic that is excluded, of course, is the *mean*, since tests are intended to discriminate among individuals, and as different subpopulations are comprised of individuals, there is no reason, in principle, why the mean test scores of the subpopulations should not also differ reliably if the subpopulations are not

assumed to be perfectly random samples of the total population. Every failure to reject the null hypothesis (i.e., that the subpopulations do not differ significantly on some essential psychometric characteristic of the test, such as reliability, validity, item characteristic curves, factor structures, etc.) strengthens the presumption that the test is not culturally biased for the subpopulations in question. Sibling data on a battery of tests provide an additional means for testing this null hypothesis.

This use of sibling data is illustrated by measurements obtained on pairs of siblings from 1,495 white families and 901 black families in grades 2 to 6 (ages of about 7 to 12 years) in California schools. In all cases, the pair of siblings in each family nearest in age and enrolled in grades 2 to 6 were selected for this study. In addition to measurements of height and weight, scores on the following tests were obtained:

Memory: a composite score based on three highly intercorrelated tests of rote memory involving immediate recall, delayed recall, and learning through repetition.

Figure Copying: copying ten geometric forms of increasing difficulty (complexity).

Pictorial IQ: Lorge-Thorndike Primary Level IQ test, a nonreading test employing pictorial materials and orally group-administered.

Nonverbal IQ: Lorge-Thorndike Nonverbal, a completely nonreading, nonverbal test employing pictorial and abstract figural material.

Verbal IQ: Lorge-Thorndike Verbal, a verbal test consisting of word similarities, opposites, verbal analogies, verbal reasoning, and so forth.

Vocabulary: Word Meaning subtest of the Stanford Achievement Test.

Reading Comprehension: Paragraph Meaning subtest of the Stanford Achievement Test.

Age Standardization: Raw scores on each of the above tests, as well as the height and weight measurements, were age-standardized (i.e., normalized standard scores within 6-month intervals) on the data of an entire school district with approximately 8,000 pupils, 60 percent whites and 40 percent blacks. Age-standardization was done separately for blacks and whites. The adequacy of the age-standardization was tested by the sibling method described under section 2. The shrunken multiple R between powers of sibling age differences and sibling standardized score differences was not significantly greater than zero for any of the measurements for either whites or blacks. In other words, age variance in scores was effectively removed by the standardization procedure.

Interval Scale. The age-standardized scores on each of the tests was checked for interval scale within each racial group by the method described in section 3. All of the tests met the criterion of interval scale for both racial groups.

Height and Weight. The BF and WF correlations between height and

weight (correlations corrected for attenuation in parentheses), and the sibling correlations, $r_{HH'}$ and $r_{WW'}$, are shown in Table II. Height and weight are obviously *intrinsically* correlated—an example of correlation through a common factor, viz., general body size.

Height and weight show only two significant ($p < .05$) WF correlations with the seven mental tests for both whites and blacks, while there are 12 significant BF correlations with the physical measurements, suggesting that the physical measurements are not intrinsically correlated with mental ability. The average WF and BF correlations between the physical and mental measurements are +.02 (ns) and +.10 ($p < .01$), respectively.

Factor Analysis of BF and WF Correlation Matrices

The BF and WF matrices of corrected correlations among the seven tests were factor analyzed separately, for whites and blacks. The general factor common to all of the tests is represented by the first principal component, which was extracted from the BF and WF correlation matrices for whites and blacks.

If the tests' intercorrelations are intrinsic rather than extrinsic, the same general factor *g* should appear in both the BF and WF matrices, which would indicate that test score differences between siblings reflect the same general factor as test score differences between children from different families. And, if the tests are not culturally biased, one should expect to find the same *g* factor in whites and blacks, for both BF and WF correlations.

Table III shows the loadings on the first principal component extracted from the four correlation matrices. It can be seen that the four factors are highly similar. The size of the loadings are generally larger for the BF than the WF first principal component, which should be expected if there is assortative mating for *g*, since all of the population variance attributable to assortative mating is BF variance. Also, there is undoubtedly some BF environmental variance, which appears to be similar in nature to WF environmental variance, considering the similarity between the BF and WF *g* factors shown in Table III.

A quantitative index, ranging from −1 to +1, commonly used to measure the degree of similarity between factors extracted from the same set of

TABLE II

BF and WF and Sibling Correlations of Height and Weight in White and Black Samples

Type of Correlation		White	Black
Between Family	r_{HW}	*.73* (*.75*)	*.79* (*.81*)
Within Family	r_{HW}	.68 (.71)	.70 (.74)
Between Siblings	$r_{HH'}$	.42 (.43)	.45 (.46)
Between Siblings	$r_{WW'}$	.38 (.39)	.37 (.38)

Note. Correlations in parentheses are corrected for attenuation.

variables in different samples is the *congruence coefficient* r_c (Cattell, 1978, pp. 251–255):

$$r_c = \sum b_i b_j / (\sum b_i^2 \sum b_j^2)^{1/2}, \tag{10}$$

where b_i and b_j are the factor loadings on the same tests in groups i and j. The congruence coefficients among the four g factors are shown in Table IV. They are all very high and do not differ significantly, indicating that this battery of tests measures the same g factor both *within* and *between* families, and in both whites and blacks in this California school population. Thus, the intercorrelations among the tests are mainly *intrinsic* correlations in both racial populations, and, so too, the g factor (first principal component) common to all of the tests is an intrinsic factor, as would be also individual factor scores derived therefrom. Either there are no cultural differences between the groups or whatever cultural differences may exist do not significantly alter the character of the general factor that is common to these diverse tests.

TABLE III

Loadings on the First Principal Component (Corrected for Attenuation) of the Between- and Within-Family Correlations for Whites and Blacks

Test	White		Black	
	Between	Within	Between	Within
Memory	.585	.285	.477	.398
Figure Copying	.492	.350	.645	.316
Pictorial IQ	.950	.789	.847	.945
Nonverbal IQ	.904	.892	.932	.871
Verbal IQ	.952	.981	.958	.924
Vocabulary	.825	.713	.698	.657
Reading Comprehension	.878	.737	.796	.688
Percent of Variance	66.6	51.9	61.0	52.4

TABLE IV

Congruence Coefficients Between First Principal Components of Whites (W) and Blacks (B), Between (b) and Within (w) Families

Principal Component	(2) W w-f	(3) B b-f	(4) B w-f
(1) White between-families	.987	.993	.991
(2) White within-families		.986	.993
(3) Black between-families			.985
(4) Black within-families			

REFERENCES

CATTELL, R. B. *The scientific use of factor analysis in behavioral and life sciences.* New York: Plenum, 1978.

CAVALLI-SFORZA, L. L., & BODMER, W. F. *The genetics of human populations.* San Francisco: W. H. Freeman, 1971.

ERLENMEYER-KIMLING, L., & JARVIK, L. F. Genetics and intelligence: A review. *Science,* 1963, *142,* 1477–1479.

FISHER, R. A. *Statistical methods for research workers* (14th ed.). New York: Hafner (Macmillan), 1970.

HAGGARD, E. A. *Intraclass correlation and the analysis of variance.* New York: Dryden, 1958.

HIRSCHI, T., & HINDELANG, M. J. Intelligence and delinquency: A revisionist review. *American Sociological Review,* 1977, *42,* 571–587.

JENSEN, A. R., *Educability and group differences.* New York: Harper Row, 1973.

JENSEN, A. R. Cumulative deficit: A testable hypothesis? *Developmental Psychology,* 1974, *10,* 996–1019.

JENSEN, A. R. Cumulative deficit in IQ of blacks in the rural South. *Developmental Psychology,* 1977, *13,* 184–191.

JENSEN, A. R. Genetic and behavioral effects of nonrandom mating. In R. T. Osborne, C. E. Noble, & N. Weyl (Eds.) *Human variation: Psychology of age, race, and sex.* New York: Academic Press, 1978.

JENSEN, A. R. *g*: Outmoded theory or unconquered frontier? *Creative Science and Technology,* 1979, *2,* 16–29.

LAYCOCK, F., & TAYLOR, J. S. Physiques of gifted children and their less gifted siblings. *Child Development,* 1964, *35,* 63–74.

OSBORNE, R. T., & SUDDICK, D. E. Blood type gene frequency and mental ability. *Psychological Reports,* 1971, *29,* 1,243–1,249.

VANVALEN, L. Brain size and intelligence in man. *American Journal of Physical Anthropology,* 1974, *40,* 417–423.

AUTHOR

ARTHUR R. JENSEN, Professor of Educational Psychology, School of Education, University of California, Berkeley, California 94720. *Specialization:* Individual differences; psychometrics; behavior genetics.