(Reprinted from Nature, Vol. 240, No. 5376, pp. 84–88, November 10, 1972)

Insignificance of Evidence for Differences in Heritability of IQ between Races and Social Classes

L. J. EAVES & J. L. JINKS

Department of Genetics, University of Birmingham, Birmingham B15 2TT

Evidence previously analysed is insufficient to support the conclusions drawn.

DURING the last few years, Jinks, Fulker and Eaves[1–6] have systematically reanalysed many of the available data on IQ and from a combination of this experience, biometrical model-building and computer simulations we have defined both the qualitative and quantitative minimal requirements for such data if they are to yield estimates of heritability and of the genetical, environmental and interactive components of variation. We have also described kinds of data and laid down guidelines for the future collection of data that would be adequate to answer the kinds of question that have been posed but so far inadequately answered.

Dr Scarr-Salapatek[7] has attempted to go beyond what we have shown to be possible with the minimal set of data we considered, doing so on the basis of analyses of data which fall short of this minimal set in both quality and quantity. It is necessary, therefore, to examine the consequences of doing so.

Qualitative Inadequacies

Qualitatively, the minimal set of data considered by Jinks and Fulker[1] consists of a number of pairs of monozygotic (MZ) and dizygotic (DZ) twins, the individuals in each pair having been raised together. Such data provide an estimate of the ratio of genetical variation within families (pairs) to the total variation arising from all sources within families[1]—the H statistic of Holzinger. This statistic is not a heritability estimate in any meaningful sense as it omits all information about the genetical, environmental and interactive sources of variation that arise between different families (pairs). It is an estimate of broad heritability only where the ratio of genetical to all sources of variation is the same both within and between families. In addition, the minimal set of data also provides a test for the presence of interactions and correlations of the genotype with the within family environment and interactions of the environmental components of variation within and between families. Such data, however, will not provide estimates of the four basic components of the total variation, namely, the genetical and environmental variation within and between families, that is, the G_1, E_1 and G_2, E_2 of Jinks and Fulker which are directly relatable to the σ^2_{wg}, σ^2_{we} and σ^2_{bg} and σ^2_{be} of Scarr-Salapatek.

The data presented by Dr Scarr-Salapatek fall short of this minimal set in that there is no complete classification of twin pairs into monozygotic and dizygotic. They are classified into twins of unlike sex that must be dizygotic in origin and twins of like sex that may be either monozygotic or dizygotic. With a notional partitioning of the twins of like sex into proportions that are monozygotic and dizygotic in origin, of the kind used by Scarr-Salapatek, the data become equivalent to the minimal set in one respect but fall short in all others. They provide an estimate of Holzinger's H statistic, but with a larger standard error, and no test for genotype-environmental interactions or correlations. In relating Scarr-Salapatek's derivation of the H statistic (her "restricted heritability" h^2_r) to that of Jinks and Fulker[1] and Eaves[5] it should be noted that the σ^2s of Scarr-Salapatek are not the variance components of the conventional analysis of variance but are the mean squares of the latter.

From the estimate of the H statistic and the corresponding total variance Scarr-Salapatek proceeds to estimate the genetical and environmental components of the variances within and between families. With only the equivalent of the minimal set of data this procedure is not possible without making assumptions[1]. The nature of these assumptions can be seen from the simplest of all models (which assumes random mating and no genotype-environmental interactions or correlations) in which G_1, E_1 and G_2, E_2 represent the genetical and environmental components of variation within and between families as follows (see Scarr-Salapatek, Table 10):

Component	Within family	Between family	Row total
Genetical	G_1	G_2	G_1+G_2
Environmental	E_1	E_2	E_1+E_2
Column total	G_1+E_1	G_2+E_2	$G_1+G_2+E_1+E_2=$ total variance (V_T)

Because

$$\frac{\text{Row 1 total}}{\text{Total variance}} = \frac{G_1+G_2}{G_1+G_2+E_1+E_2}$$

is the true broad heritability, h^2_b, and

$$\frac{\text{Column 2 total}}{\text{Total variance}} = \frac{G_2+E_2}{G_1+G_2+E_1+E_2}$$

is the intraclass correlation, r_{dz}, for dizygotic twins, the row totals equal $h^2_b V_T$ and $(1-h^2_b)V_T$ and the column totals $(1-r_{dz})V_T$ and $r_{dz}V_T$, respectively. From Scarr-Salapatek's data we can estimate only $h^2_r = G_1/(G_1+E_1)$ and to equate this statistic to h^2_b we must assume that $G_1/G_2 = E_1/E_2$. This is also a necessary assumption for the next step in Scarr-Salapatek's analysis which is the estimation of G_1, G_2, E_1 and E_2 from the row and column totals.

The relative magnitudes of G_1 and G_2 depend on the kinds of gene action underlying the variation and the mating structure of the population[1]. In the absence of both dominance and assortative mating $G_1 = G_2$, with dominance alone $G_1 > G_2$

and with assortive mating alone $G_1 < G_2$. Both dominance and assortative mating are known to occur for IQ[1,6,8,9] and since they affect the relative magnitudes of G_1 and G_2 in opposite directions we neither expect nor find large differences between them.

The relative magnitudes of E_1 and E_2 cannot be predicted from any *a priori* model; they can only be established empirically by observation. The minimal set of data which allows the estimation of E_1, if we assume the present model, cannot provide a direct estimate of E_2. Thus, the assumption that $G_1/G_2 = E_1/E_2$, that underlies the analyses and interpretations of Scarr-Salapatek, is neither testable from the data she provides nor can it be justified on theoretical grounds. These arguments are, of course, made more complex if we attempt, as does Dr Scarr-Salapatek, to correct h_r^2 and the components of the total variation for the effects of assortative mating (her h_a^2) but this extension does not invalidate the principle we have sought to illustrate by reference to the simpler situation, namely, that her analysis involves untestable assumptions about the relative magnitudes of the genetical and environmental components.

Quantitative Inadequacies

Having commented upon the limitations imposed on the analysis and interpretation arising from the qualitative aspects of the data, we can now turn our attention to the limitations that arise from the quantitative aspects which depend on the number of twin pairs that fall within each of the racial, sex and socio-economic sub-groups. While it is the qualitative properties of the data that determine the kinds of analyses and conclusions that can be validly applied, it is the quantitative properties that determine the standard errors of the estimates, their significance levels and hence the confidence that can be placed on the conclusions.

Dr Scarr-Salapatek provides no errors for her estimates of "heritability" (H statistics) and she compares and interprets these estimates with no regard to their likely errors. Elsewhere it has been argued that even data which are qualitatively adequate will not yield convincing and significant results unless sample sizes are much larger than those employed in this study[5]. It is no surprise, therefore, that when we attempt to derive standard errors for some of the comparisons made by Dr Scarr-Salapatek we find that little confidence can be placed in individual "heritability" estimates and even less upon comparisons between them.

In deriving conclusions from the raw correlations, Dr Scarr-Salapatek combines correlations firstly to estimate the intraclass correlation for monozygotic twins, (r_{mz}), secondly to estimate the "heritability", (h_r^2), and finally to compare "heritability" estimates from different subpopulations. We shall show that the tests of significance, which should be applied before strong conclusions are claimed, are practically powerless with the sample sizes used in her study. Indeed, even gross effects could not be detected.

Consequences of Indirect Estimation of r_{mz}

The correlation between monozygotic twins is estimated from the z values obtained for same-sex (SS) and opposite sex (OS) pairs. If the proportion of OS pairs in the population is p, then:

$$z_{ss} = \frac{pz_{os} + (1-2p)z_{mz}}{1-p}$$

giving:

$$z_{mz} = \frac{(1-p)z_{ss} - pz_{os}}{1-2p}$$

Dr Scarr-Salapatek uses r instead of z in connexion with these formulae[7] (p. 1287), although her estimates of the MZ correlations are, in fact, correctly based on the z's. The variance of z_{mz} is given by:

$$\sigma^2_{z_{mz}} = \left(\frac{1-p}{1-2p}\right)^2 \sigma^2_{z_{ss}} + \left(\frac{p}{1-2p}\right)^2 \sigma^2_{z_{os}}$$

assuming p to be known exactly.

For whites p is given (p. 1288) as 0.3, which yields

$$\sigma^2_{z_{mz}} = 3.0625\sigma^2_{z_{ss}} + 0.5625\sigma^2_{z_{os}}$$

and for blacks ($p=0.34$)

$$\sigma^2_{z_{mz}} = 4.2539\sigma^2_{z_{ss}} + 1.1289\sigma^2_{z_{os}}$$

These values of σ_z^2 are inversely related to the sample sizes only. For a given number of SS and OS pairs, it is a simple matter to calculate $\sigma^2_{z_{mz}}$ since $\sigma_z^2 = 1/(N-3) \simeq 1/N$ for large samples, where N=number of pairs. In Dr Scarr-Salapatek's samples, SS pairs are approximately twice as frequent as OS pairs, so $\sigma^2_{z_{ss}} = 1/2N_{os} = \frac{1}{2}\sigma^2_{z_{os}}$

where N_{os} is the number of OS pairs in the sample.

Thus,

$\sigma^2_{z_{mz}} = 2.09\sigma^2_{z_{os}}$ for whites and

$\sigma^2_{z_{mz}} = 3.26\sigma^2_{z_{os}}$ for blacks.

The standard error of the restricted heritability (h_r^2) cannot be estimated directly for reasons already stated, but it is arguably pointless to produce such an estimate unless the difference $z_{mz} - z_{dz}$ is itself significant, because this difference is the numerator in the estimation of h_r^2.

The variance of the difference is:

$$\begin{aligned}\sigma_d^2 &= \sigma^2_{z_{mz}} + \sigma^2_{z_{dz}} \\ &= \sigma^2_{z_{mz}} + \sigma^2_{z_{os}} \\ &= 3.09\sigma^2_{z_{os}} \text{ for whites}\end{aligned}$$

and $4.26\sigma^2_{z_{os}}$ for blacks.

given samples of the same proportions as before. This estimate of σ_d^2 applies only when the indirect method of estimating r_{mz} is used. Given accurate zygosity determination on the other hand, and assuming equal numbers of monozygotic and dizygotic twins:

$$\begin{aligned}\sigma_d^2 &= \sigma^2_{z_{dz}} + \sigma^2_{z_{mz}} \\ &= 2\sigma^2_{z_{os}}\end{aligned}$$

Thus, a sample of N MZ pairs and N DZ pairs gives a value of σ_d^2 which is approximately half that obtained for a sample of N OS pairs and $2N$ SS pairs of unknown zygosity. If zygosity determination is not undertaken, therefore, the size of the experiment has to be increased by a factor of, approximately, three to avoid loss of power in testing for a genetical component. This is a very damaging consequence of the indirect method of estimating the correlation between MZ twins which it may be difficult to justify on economic grounds.

Power of the Test for a Genetical Component

For a given true heritability, with certain assumptions about gene action, the mating system and environmental variation, expected values of the correlations between MZ and DZ twins can be derived. Knowledge of the standard error of the difference $z_{mz} - z_{dz}$ and the expected value of the difference enables the power of the test to be calculated for samples of a given size. That is, we can calculate for a given sample structure the probability of correctly rejecting the null hypothesis that there is no genetical component of variation. If this probability is low then the test is poor since the null hypothesis will be generally retained even though false (type II error).

There is a prior expectation that $z_{mz} \geq z_{dz}$, so the test of the difference $d = z_{mz} - z_{dz}$ is a one-tail test. That is, if

$$c = (z_{mz} - z_{dz})/\sigma_d \geq 1.65$$

we reject at the 5% level the null hypothesis that there is no heritable variation. For a given expected $z_{mz} - z_{dz}$, which depends upon the true heritability, and for a given σ_d, which depends upon the sample size, the expected value of c can be calculated, c_e. The power of the test is then the area under a normal curve with zero mean and unit variance between the limits $(1.65 - c_e)$ and infinity.

If 60% of the variation is genetically determined and there are no common environmental effects, the expected value of r_{mz} would be 0.6. This is approximately the mean value of r_{mz} given in Dr Scarr-Salapatek's study, and is an upper limit to the true broad heritability of the trait. If, further, there is no dominance, the expected value of r_{dz} will be 0.3, providing mating is at random. Under conditions of assortative mating the DZ correlation will be higher, being 0.45 if there is a correlation of 0.5 between the additive genetical deviations of spouses.

The expected value of $z_{mz} - z_{dz}$ will then be 0.3836 in a randomly mating population and 0.2084 under assortative mating of the kind just defined.

Consider the sample of upper socio-economic status (SES) whites, consisting of 70 OS pairs[7] (Table 8, page 1291). Assume, for approximation, that the number of SS pairs, actually 155, is exactly twice that of OS pairs, so that $\sigma_d^2 = 3.09\sigma_{z_{os}}^2$ as above.

$$\text{Now } \sigma_{z_{os}}^2 = 1/70$$
$$= 0.014286$$

$$\text{so that } \sigma_d^2 = 0.044143$$

$$\text{and } \sigma_d = 0.2101.$$

For the randomly mating population the expected value of c is thus

$$c_e = 0.3836/0.2101$$
$$= 1.8258, \text{ when } h_b^2 = 0.6.$$

The power of the test, α, is thus the area under a normal curve having zero mean and unit variance, between the limits -0.18 and infinity. In this case α can be found from tables to be 0.57. That is, a significant genetical component of variation will only be detected in randomly mating populations in 57% of all possible samples of this size, even when the broad heritability is as high as 0.6. Under conditions of assortative mating a similar calculation shows that samples of this size would only produce a significant genetical component in 25% of studies. Table 1 gives the power of the test for the four separate subclasses of Scarr-Salapatek's study by race and SES, and the value of α for tests for each race separately after pooling over social classes. The sample sizes hardly provide powerful tests of a genetical component when the subgroups are considered separately, and do not provide a very rigorous test even when a pooled "heritability" estimate is obtained for each race. It is noticeable that a moderate degree of assortative mating reduces the power of the tests to values which would inevitably provide non-significant estimates more often than not. To provide more convincing tests (say, $\alpha = 0.95$), between 800 and 1,000 pairs are needed for randomly mating populations, and between 2,000 and 3,500 pairs are needed for assortatively mating populations, depending on race. If we remove the simplifying assumptions of no dominance or E_2 we find that the presence of either will tend to improve the power of the test, dominance by reducing r_{dz} relative to r_{mz}, and E_2 by increasing the overall correlation between relatives and thus, on the transformed scale, increasing the difference $z_{mz} - z_{dz}$.

Comparing "Heritabilities"

The conclusions reached so far relate only to the existence or otherwise of a genetical component of variation. We have seen that even a relatively large genetical component, corresponding to a true broad heritability (h_b^2) of 0.60, can only be detected unreliably with samples of this size. Dr Scarr-Salapatek's conclusions, however, are based on the comparison of estimates of h_r^2 for different subpopulations, so we must enquire to what extent statistical unreliability is increased by attempting to draw comparative conclusions about different groups of individuals. We will concern ourselves only with a comparison between races.

The null hypothesis, that there is no racial difference in "heritability", is only rejected if the comparison $k = (z_{mz} - z_{dz})_{white} (z_{mz} - z_{dz})_{black}$ differs significantly from zero.

The variance of this comparison is

$$\sigma_k^2 = \sigma_{d,\ white}^2 + \sigma_{d,\ black}^2.$$

For samples in which like-sex twins are twice as frequent as unlike sex pairs

$$\sigma_k^2 = 3.09\sigma_{z_{os},\ white}^2 + 4.26\sigma_{z_{os},\ black}^2.$$

There is no prior expectation about the direction of the difference so the null hypothesis will only be rejected at the 5% level if $k/\sigma_k \geq 1.96$.

With the sample sizes used in the study, $\sigma_{z_{os},\ white}^2 = 0.011628$ and $\sigma_{z_{os},\ black}^2 = 0.005917$, so that

$$\sigma_k^2 = 0.061137$$

and $\sigma_k = 0.2473$.

In an extreme case, where the true heritability in one population is 0.6, and there is no heritable variation in the other:

$$c_e = 0.3836/0.2473 = 1.55$$

under conditions of random mating. The power of the test is thus the area under a normal curve of unit variance between the limits $(1.96 - 1.55)$ and infinity. That is, the power of the test (α) is 0.34. Thus, even in this extreme case, we shall find, more often than not, that there is no significant difference in the genetical structure of the two populations with the sample sizes in Dr Scarr-Salapatek's study. With equal numbers of black and white pairs, nearly 4,000 pairs would be needed altogether to be 95% certain of detecting a difference of the grossest kind between the heritabilities of a trait in the two populations. If the difference is less marked, say $h_b^2 = 0.3$ in one population and 0.6 in the other, over 3,000 pairs are required for the power of such a test to be even 0.5, and upwards of 11,000 pairs would be needed before we could be 95% certain of detecting a difference between the two heritabilities. On purely theoretical grounds, therefore, we suggest that this particular experimental design, with the small samples available, could not be expected to lead to the conclusions which were drawn and indeed could only be drawn from it by omitting proper tests of significance.

Table 1 The Power of the Test for a Genetical Component of Variation

	Subpopulation					
	Black			White		
	Low SES	High SES	Pooled	Low SES	High SES	Pooled
Total sample size (N)	321	186	507	48	210	258
Power of test (α)						
Random mating	0.61	0.43	0.78	0.22	0.57	0.64
Assortative mating	0.27	0.20	0.37	0.14	0.25	0.29

Sample sizes approximately equal to those in Scarr-Salapatek[7]. A broad heritability (h_b^2) of 0.6 is assumed, and values are tabulated for randomly mating and assortatively mating populations. For simplicity no dominance or E_2 has been assumed.

The Simplest Model

We reanalysed Dr Scarr-Salapatek's data using a more rigorous approach to see whether any statistical significance could be attached to the strong conclusions she draws from the tabulated correlations. We describe an analysis of variation of the z values for the cells of a table of correlations for the two types of twin in each race and SES combination (derived from Scarr-Salapatek's Tables 6 and 7, p. 1291). We observe, firstly, that such a detailed analysis is not strictly justified by the data because the eight raw correlations are homogeneous for the non-verbal scores ($\chi^2_{(7)} = 5.63$, $50\% < P < 75\%$) and they are barely heterogeneous for the verbal scores ($\chi^2_{(7)} = 15.63$, $2\frac{1}{2}\% < P < 5\%$). This means that all the correlations given by Dr Scarr-Salapatek are really nothing more than estimates of the same population value of the correlation between twins, irrespective of their classification as SS or OS. We give, however, a more detailed analysis of the correlations for the verbal scores because of the slight indication of heterogeneity. The variation in the z's for the eight correlations can be predicted by the linear model given in Table 2. The model includes, besides the overall mean value of z, the effects due to race, SES, and the difference between SS and OS twins. Of particular interest in the light of Dr Scarr-Salapatek's analysis, however, is the possibility of attaching tests of significance to the first order interaction between the SS/OS dichotomy and social class and that between SS/OS and race which provide the crucial tests of differences in heritability between races and social classes.

Because the z values are based on different numbers of observations they do not have the same variance so the estimated components of the linear model are not orthogonal. The method of weighted least squares, however, yields maximum likelihood estimates of the effects and gives their variance-covariance matrix. These estimates are given in Table 3 with their standard errors derived from the diagonal elements of the variance–covariance matrix. All the estimates have the same standard error since every z enters into each comparison. The fact that the only significant effect is the overall mean suggests that the slight heterogeneity of the z values cannot be assigned to any particular cause.

We find, first, that there is no significant overall difference between the correlations for SS and OS twins. This implies that the data cannot even support the well-established conclusion that there is a genetical component of individual differences in intelligence. We find further that the interactions of the SS/OS difference with race and SES are not significant. This confirms that there is no evidence that the size of any heritable component depends on race or social advantage. This finding contradicts the main conclusion of Dr Scarr-Salapatek's analysis which is based on a comparison of the numerical values of the correlations.

As there is no detectable heritable component, we cannot, on the basis of this study, suppose that the similarity between twins is due to anything other than common environmental effects. Such a conclusion is clearly inconsistent with other, more secure, evidence on this matter[1,6,9–12]. The fact that the overall correlations for both types of twin depend neither on race nor on socio-economic status indicates that there is no difference in the magnitude of a common environmental component between the races or the two social groupings. Furthermore, the absence of a race × socio-economic status interaction implies that the magnitude of any common environmental effect does not depend on the joint effects of race and social class.

The only tenable conclusion to be drawn from the data is that there is a highly significant correlation between twins of all kinds for verbal IQ ($z = 0.597$, $P = 0.001$, $r = 0.54$). We are in no position to decide the cause of such similarity. There is no evidence that it has a genetical basis as far as this study goes, but as we have shown above, the likelihood of detecting such an effect with this experimental design and with these samples is very small. There is certainly no evidence in Scarr-Salapatek's studies that the proportion of genetical variation in either verbal or non-verbal IQ depends on race or social class. In view of this conclusion, and having regard to the general absence of genotype-environmental interactions for IQ[1,13,14], there is little justification for detailed consideration of the particular models suggested by Dr Scarr-Salapatek.

We thank Professor P. L. Broadhurst for helpful comments on the manuscript. This work is part of a research project in psychogenetics supported by the Medical Research Council.

Received August 3, 1972.

Table 2 Linear Model for Predicting the Observed Degree of Similarity between Twins (Measured by z) in Terms of Race, Social Class and Concordance for Sex

	Black				White			
	Low SES		High SES		Low SES		High SES	
	OS	SS	OS	SS	OS	SS	OS	SS
Mean	1	1	1	1	1	1	1	1
Race	1	1	1	1	−1	−1	−1	−1
Socio-economic status (SES)	1	1	−1	−1	1	1	−1	−1
Same sex *v.* opposite sex pairs (SS/OS)	1	−1	1	−1	1	−1	1	−1
Race × SS/OS	1	−1	1	−1	−1	1	−1	1
Race × SES	1	1	−1	−1	−1	−1	1	1
SES × SS/OS	1	−1	−1	1	1	−1	−1	1
Race × SES × SS/OS	1	−1	−1	1	−1	1	1	−1

Table 3 Effects Contributing to Variation in the Similarity Between Twins for Verbal IQ

Effect	Estimate
Mean	0.597*
Race	−0.048 NS
SES	0.052 NS
SS/OS	−0.069 NS
Race × SS/OS	0.025 NS
Race × SES	0.008 NS
SES × SS/OS	−0.053 NS
Race × SES × SS/OS	0.018 NS

The estimates are obtained by weighted least squares from the observed values of z. The standard error of every estimate is 0.051.

*=Significant at the 0.1% level.

NS=Not significant at the 5% level.

1 Jinks, J. L., and Fulker, D. W., *Psychol. Bull.*, **73**, 311 (1970).
2 Fulker, D. W., *Symposium on Methodology in Human Behaviour Genetics, 4th Int. Cong. Hum. Genet.* (1971).
3 Eaves, L. J., *Brit. J. Math. Statist. Psychol.*, **22**, 131 (1969).
4 Eaves, L. J., *Brit. J. Math. Statist. Psychol.*, **23**, 189 (1970).
5 Eaves, L. J., *Psychol. Bull.*, **77**, 144 (1972).
6 Eaves, L. J., *Heredity* (in the press).
7 Scarr-Salapatek, S., *Science*, **174**, 1285 (1971).
8 Reed, E. W., and Reed, S. C., *Mental Retardation: A Family Study*, 57 (Saunders, Philadelphia, 1965).
9 Burt, C., *Brit. J. Psychol.*, **57**, 137 (1966).
10 Jensen, A. R., *Harv. Educ. Rev.*, **39**, 1 (1969).
11 Erlenmeyer-Kimling, L., and Jarvik, L. F., *Science*, **142**, 1477 (1963).
12 Nichols, R. C., in *Methods and Goals in Human Behaviour Genetics* 231 (edit. by Vandenberg, S. G.) (Academic Press, New York, 1965).
13 Jensen, A. R., *Behavior Genetics*, **1**, 133 (1970).
14 Eaves, L. J., PhD thesis, University of Birmingham (1970).

Printed in Great Britain by Flarepath Printers Ltd., St. Albans, Herts.