Am J Hum Genet 35:1179–1189, 1983

Errors of Inference in the Detection of Major Gene Effects on Psychological Test Scores

LINDON J. EAVES[1]

SUMMARY

Computer simulation methods were employed to generate abilities of 10 sets of 250 nuclear families, each comprising a pair of randomly mated parents and two children. It was assumed that the distribution of abilities in the population was normal and caused entirely by additive polygenic effects. A simulated psychological test was administered to each sample to generate test scores for each subject. A different test, consisting of 40 items of varying difficulty and discriminating power, was used in each sample. The "mixed model," specifying a single major gene with polygenic and environmental background variation, was tested for each data set. Likelihood ratios were computed to test for the contribution of a major locus and its conformity to Mendelian segregation. Only one out of 10 samples was consistent with pure multifactorial inheritance. Of the remaining nine samples, four showed non-Mendelian segregation and five were consistent with current statistical criteria for establishing the contribution of a major gene to variation in psychological test scores. This high frequency of false conclusions suggests that the naïve application of such methods to behavioral data is often likely to be misleading. Raw test scores alone are not sufficient to test the mixed model. The development of tractable models for behavioral traits requires the responses of subjects to individual items.

INTRODUCTION

The formulation of the "mixed model" for continuous variation in human populations has been a significant recent advance in human quantitative genetics

Received January 25, 1983; revised March 21, 1983.

This work was supported in part by grant 1R01-GM30250-01 from the National Institutes of Health. This is paper no. 170 from the Department of Human Genetics of Virginia Commonwealth University.

[1] Department of Human Genetics, Medical College of Virginia, Box 33, MCV Station, Richmond, VA 23233.

(e.g., [1, 2]). The model partitions the variation for a quantitative trait into three sources: the effects of a single gene of large effect; residual additive effects of polygenic loci; and the independent random effects of the environment. Numerous applications of the mixed model have now been published (e.g., [3]). In the case of hypercholesterolemia, at least, the evidence for conformity with the mixed model is overwhelming [2].

The robustness of genetic conclusions based on the statistical model alone have been examined from several viewpoints (e.g., [4, 5]). Such studies have suggested that nonnormality in the data, chiefly skewness, could lead to spurious support for a major gene since the likelihood under the classical polygenic model is computed on the assumption of multivariate normality of the trait in the sample pedigrees. In light of these considerations, Elston [2] advised caution in inferring major locus segregation in advance of evidence from genetic linkage studies.

Any problems that apply to biochemical or physiological measures on account of scaling apply, a fortiori, to psychological tests in which the process of scaling is even more arbitrary. Even though the theory underlying the construction of a particular test may require that abilities be normally distributed, the items chosen to estimate abilities are normally discontinuous, vary in their difficulty and discriminating power, and generate ability scores whose distribution can be engineered to almost any form at the behest of the psychometrician. Although it is possible to establish criteria for the construction of tests and the estimation of scores that maximize the agreement between the observed and expected distribution of abilities [6], such sophistication is rarely realized in practice even in the domain of ability measurement. A judicious choice of test items may, in theory at least, generate test scores that have uniform error variance over a wide range of ability or that maximize test-discriminating power over a particular restricted range of interest, as in the case of a test designed to select candidates for graduate study. Conversely, an incautious selection of items may produce test scores that appear to group a population into classes that could subsequently be confused with genetically distinct categories.

An intimate connection is, therefore, to be expected between the detailed outcome of the genetic analysis of psychological test scores and the properties of the items selected to measure the underlying normal abilities. As long as the genetic analysis of psychological data concentrates on estimating genetic and environmental components of variance (or path coefficients) from summarized kinship data, such problems of measurement are unlikely to lead to gross errors of inference. The detection of major gene effects, however, relies on far more subtle properties of the multivariate distribution of traits in families and is much more likely to be sensitive to the problems of scaling associated with psychological test construction.

This study tries to quantify errors of inference when the "mixed model" is applied to psychological test data. Statistical tests for the effects of a major gene are applied to simulated scores generated by giving a series of ideal simulated psychological tests to subjects whose variation in underlying ability is due solely to polygenic inheritance.

METHODS

Simulating Latent Abilities

Ten sets of 250 randomly sampled nuclear families were simulated on the assumption that the underlying distribution of abilities in the population of families was multivariate normal. Each family consisted of parents and two children and was represented by a single four-variate normal deviate. The population covariance matrix of parents and children was assumed to be:

Mother.....	1	0	0.5	0.5
Father.....	0	1	0.5	0.5
First child.....	0.5	0.5	1	0.5
Second child.....	0.5	0.5	0.5	1

The covariance matrix is implied if the underlying abilities are completely genetically determined, gene action is additive, and mating is random. The abilities were assumed to have a mean of zero in all family members. Assortative mating for intelligence test scores is firmly established, but a satisfactory form of the mixed model for assortative mating has yet to be published. The assumption of additive gene action has also been questioned but detailed understanding of the transmission of intelligence still awaits more critical designs for the resolution of the mechanisms of mate selection, cultural inheritance, and genetic nonadditivity. Although the model assumes that the "true" abilities of the subjects are genetic, the testing process will introduce sampling variation into the measurement of ability and will result in correlations less than .5 for the actual test scores of first-degree relatives.

Simulating Psychological Tests

Ten "tests" were simulated, one for each sample of families. One of the tests was then "administered" to each of the 1,000 subjects in a given population. Every individual in a given sample received the same test. This strategy of test administration comes closest to common practice: each investigator uses his or her own test on his or her population. Each test was constructed by sampling items at random from a pool of items with known distribution of difficulty and discriminating power.

Following Birnbaum [7], we assume that the probability, P, of a correct response to a given item by a subject of ability, θ, is given by the logistic function: $P = 1/1 + e^{-a(\theta - b)}$. The parameter b is a measure of the item diffculty since the greater the value of b, the lower the probability of a correct response for all values of θ. The discriminating power of an item is given by a. Large values of a imply that the probability of a correct response changes very rapidly over a short range of ability.

Each test consisted of 40 items selected at random from an infinite population of items. The distribution of item difficulties (b) was assumed to be uniform over the ability range -3, $+3$. The distribution of difficulties, therefore, makes the tests discriminate fairly uniformly over the practical range of normal ability. Many tests of ability do not have such "ideal" properties because they are not so designed. Nevertheless, such tests are used in genetic analyses. The discriminating powers of the items (a) were assumed to be N[1.02, 0.34].

Simulating Subjects' Item Responses and Test Scores

Item responses for each subject in a sample were simulated by substituting the subject's ability (θ), and the item parameters a and b in the logistic function above. If a random

uniformly distributed number on the range 0,1 exceeded the function value, a correct response was assumed for the given item. Otherwise, the subject was scored as having answered the item incorrectly.

The test score of each subject was the number of correct responses. Although, in the ability domain, test scores are often transformed to improve normality, such transformations are rarely employed in the personality domain where simple raw test scores are used for psychological counseling and genetic analysis.

Genetic Analysis

Preliminary data summaries were computed consisting of means, standard deviations, and coefficients of skewness and kurtosis for each of the 10 samples of 1,000 individuals without regard for the family relationships. The covariance matrices between relatives were also computed for each sample.

The method of maximum likelihood was employed to compute parameter estimates and log-likelihood values for each sample under three hypotheses: (1) additive polygenic inheritance and random environmental effects; (2) a gene of large effect plus polygenic and environmental residual variation for each genotype at the major locus; and (3) a non-Mendelian major factor with polygenic and environmental background variation. Before the true mixed model is adopted for a given quantitative trait, it must be shown that the major-locus model gives a significantly greater likelihood than the classical polygenic model and that the non-Mendelian model does not improve significantly on the major-locus model. In every case, it was assumed that the polygenic effects were normally distributed and, in the case of the mixed and non-Mendelian models, that the covariance matrices of the residual effects were identical for all values of the major factor.

In practice, the polygenic model is represented by three parameters: the population mean (m); the additive genetic variance (V_A); and the random within-family environmental variance (V_E). The true (Mendelian) mixed model retains the additive genetic variance and the environmental variance as parameters describing the residual variation but requires that separate means be estimated for each of the three genotypes of the major locus. These are parameterized conveniently by a mid-homozygous value, m, an additive genetic deviation, d, and a heterozygous deviation, h [8]. The genotypic effects and the variances within each genotype are thus:

Genotype	AA	Aa	aa
Mean	$m + d$	$m + h$	$m - d$
Variance	$V_A + V_E$	$V_A + V_E$	$V_A + V_E$.

In addition, the frequency (p) of one allele at the major locus must be estimated. The mixed model thus requires the addition of three parameters to the classical model, although the setting of either p or d to zero is sufficient to reduce the mixed model to the classical form. It is, thus, not clear whether the likelihood-ratio test for the mixed model against the classical model should have 2 or 3 df. A conservative policy of basing tests on 3 df was adopted here.

Under the hypothesis of Mendelian inheritance, the probability that the three genotypes AA, Aa, and aa will each transmit the A allele are 1, .5, and 0, respectively. Go et al. [5] propose that a Mendelian hypothesis be adopted only if no further significant increase in likelihood be obtained when the estimates of transmission probabilities ($t_1 \ldots t_3$) are freed from the constraints implied by classical Mendelian inheritance. In practice, with the 10 samples in this study, complete removal of these constraints resulted in the optimization problem being so ill-conditioned that independent estimation of the three transmission probabilities was impossible. For this reason, two non-Mendelian models were fitted and the likelihoods compared with those under the Mendelian model.

In the first non-Mendelian model, the transmission probabilities for AA and aa were fixed at 1 and 0, respectively, but that for Aa was allowed to take its own value. This

leads to a likelihood-ratio test for 1 df in comparing the Mendelian and non-Mendelian versions of the mixed model. In the second model, the transmission probability of *Aa* was fixed at .5 and those of *AA* and *aa* allowed to take their own value. The likelihood ratio for this comparison with the Mendelian model has 2 df. Even with these constraints imposed, the solutions were not all uniformly easy to secure.

The generation and preliminary summary of the simulated data was conducted on the genetics department's (Medical College of Virginia) PDP 11/44 computer using a FORTRAN program. The numerical analysis required for fitting and testing the genetic models was conducted on the Virginia Commonwealth University's IBM 370/168 computer, using a double precision FORTRAN program and employing the Numerical Algorithms Group's [9] subroutine E04JBF for maximization of the likelihood function.

RESULTS

Figure 1 gives the histogram of test scores for all 1,000 subjects in the sixth simulated sample of families. As might be predicted from the uniform distribution of item difficulties, the observed distribution of test scores is unimodal and typical of many in the psychological literature. Tests that are constructed to meet different

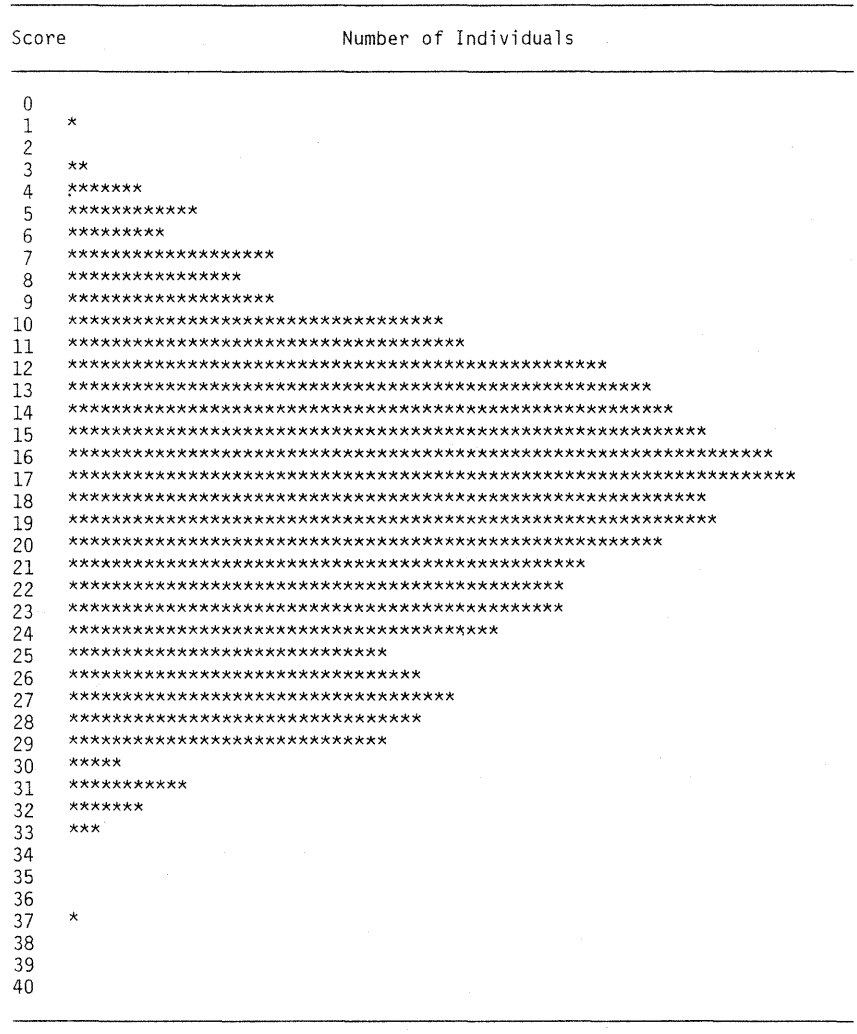

FIG. 1.—Distribution of test scores for the 1,000 subjects in data set no. 6

TABLE 1

SUMMARY STATISTICS FOR TEST SCORES IN ALL 10 SIMULATED POPULATIONS

Set no.	Mean	SD	Skewness	Kurtosis
1.....	20.4970	5.8416	0.1670	−0.2168
2.....	17.0090	5.9949	0.1916	−0.4222
3.....	17.4940	5.9016	0.1087	−0.1535
4.....	19.9650	6.6540	0.0585	−0.4650
5.....	19.1300	6.0818	0.0212	−0.4649
6.....	17.7900	6.1934	0.0882	−0.4164
7.....	23.0300	5.9110	−0.5208	−0.0588
8.....	20.5500	6.1882	−0.0853	−0.4602
9.....	17.1910	6.0195	0.3065	−0.3668
10.....	18.5260	5.9961	−0.0456	−0.2163

NOTE: The SE of the skewness and kurtosis estimates are 0.0773 and 0.1545, respectively. No. = 1,000 individuals per set.

criteria might have less ideal properties for genetic analysis. Means, standard deviations, and coefficients of skewness and kurtosis were computed for all 10 samples. These summary statistics for all 10 samples are given in table 1.

Although the underlying abilities were simulated to be normal, the resulting test scores were far from normal in eight of the 10 sets when departures from normality are judged by skewness and kurtosis. Four sets showed significant skewness when all 1,000 subjects were entered into the calculation. In two sets, the skewness was highly significant (positive in one case and negative in the other). Significant negative kurtosis was encountered in six sets. In four cases, significant kurtosis was not accompanied by significant skewness.

The means and standard deviations of latent abilities and test scores for mothers, fathers, and children in sample no. 6 are given in table 2. The matrix of the correlations between abilities and test scores of parents and children is given in table 3. Because the testing process introduces sampling variation into the measurement of ability, the correlations between the scores of relatives are lower than those between the latent abilities. The correlations between latent ability and test score may be used to predict the reliability of the 40-item test.

Table 4 gives the parameter estimates, log likelihoods, and likelihood-ratio chi squares for the four separate models fitted to the 10 data sets. In data sets 5,

TABLE 2

MEANS AND STANDARD DEVIATIONS OF SIMULATED LATENT ABILITIES AND PSYCHOLOGICAL TEST SCORES FOR 250 MOTHERS, FATHERS, AND PAIRS OF CHILDREN (DATA SET NO. 6)

	ABILITY		TEST SCORE	
INDIVIDUAL	Mean	SD	Mean	SD
Mother.....	−0.0296	0.991	17.50	6.226
Father.....	−0.0193	0.974	18.35	6.163
Child 1.....	−0.0785	1.002	17.62	6.287
Child 2.....	0.0039	0.978	17.70	6.098

TABLE 3

FAMILIAL CORRELATIONS FOR LATENT ABILITIES (UPPER TRIANGLE) AND PSYCHOLOGICAL TEST SCORES (LOWER TRIANGLE) FOR DATA SET NO. 6

	Mother	Father	Child 1	Child 2
Mother	.902	.066	.564	.551
Father	.046	.893	.469	.481
Child 1	.487	.371	.913	.425
Child 2	.488	.348	.374	.899

NOTE: The diagonal contains correlations between abilities and test scores.

6, and 10, either convergence was not achieved for the parameters of the second non-Mendelian model or trial values could not be found that gave a positive-definite estimate of the covariance matrix of family members at all stages in the optimization. Apart from these three cases, the optimization program appeared to operate satisfactorily.

Only the last data set conformed to the classical polygenic model used to generate the latent abilities of the 250 families. This data set was also one of two (sets 3 and 10) that showed no evidence of skewness of kurtosis on the univariate test. Sets 2–5 all gave a significantly greater likelihood when the mixed model was fitted but a further significant improvement was achieved when non-Mendelian transmission probabilities were estimated. In these four cases, therefore, the failure of the classical polygenic model would be ascribed correctly to nongenetic factors, either cultural transmission or problems of scaling. The remaining data sets, 1 and 6–9, all showed a marked improvement under the mixed model but did not reveal that the failure of the classical model was nongenetic. These five samples, therefore, would have led to the mistaken identification of a major locus when, in reality, the latent abilities are polygenic and multivariate normal in pedigrees. Two of the data sets (7 and 9) consistent with the mixed model were those with highly skewed scores and confirm the common finding [4, 5] that skewness in the distribution of genetic effects can often be mistaken for the effects of a major gene. The other three sets consistent with the mixed model showed no skewness but significant kurtosis in the raw scores. Either the mixed or the non-Mendelian models gave significantly improved likelihood with all samples showing significant skewness or kurtosis. A non-Mendelian model also gave a better fit to one of the two samples in which there was no skewness or kurtosis.

DISCUSSION

The most disturbing conclusion of the simulations is that such a large proportion of the genetic variance is assigned to the effects of a major locus when the mixed model is mistakenly adopted for variation in test scores. The estimates of the additive genetic variance given in table 4 under the classical polygenic model and the mixed model show that in cases where the mixed model fits best, upward of 50% of the genetic variance is mistakenly ascribed to a major gene. Such a

TABLE 4

SUMMARY STATISTICS FROM GENETIC ANALYSIS OF 10 SIMULATED DATA SETS

SET	MODEL	PARAMETER ESTIMATE									LOG(L)	χ^2	P
		V_A	V_E	m	p	d	h	t_1	t_2	t_3			
1	I.....	28.04	5.76	20.49	...	...	...	...	...	...	3072.4	...	...
	II.....	13.11	5.63	21.83	.421	5.67	−0.90	1	.5	0	3068.2	8.44	*
	III.....	13.11	5.63	21.83	.421	5.67	−0.90	1	.5	0	3068.2	...	...
	IV.....	13.09	4.59	21.84	.414	5.86	−0.72	.953	.5	.028	3068.1	0.18	...
2	I.....	26.85	8.67	17.09	...	...	...	...	...	...	3117.8	...	...
	II.....	14.85	6.72	18.16	.558	4.43	−3.25	1	.5	0	3110.3	15.08	**
	III.....	14.98	6.19	17.78	.209	3.51	4.48	1	.415	0	3107.2	6.02	*
	IV.....	9.79	6.66	19.35	.322	6.50	3.20	.776	.5	.003	3106.8	6.84	*
3	I.....	27.80	7.09	17.55	...	...	...	...	...	...	3097.8	...	...
	II.....	18.74	6.12	19.15	.245	4.33	1.67	1	.5	0	3096.1	3.42	...
	III.....	19.82	5.47	19.09	.227	4.06	2.13	1	.422	0	3095.2	1.76	...
	IV.....	19.93	0.06	18.65	.371	5.44	0.76	.5†	.5	.101	3091.0	12.30	***
4	I.....	38.47	6.19	20.02	...	...	...	...	...	...	3201.3	...	...
	II.....	21.81	5.12	20.09	.634	4.94	−3.02	1	.5	0	3197.2	8.26	*
	III.....	24.81	2.81	20.30	.256	3.95	4.42	1	.453	0	3194.2	6.00	*
	IV.....	26.76	−2.68	21.90	.330	6.47	0.85	.5†	.5	.105	3193.8	8.88	*
5	I.....	31.06	5.30	19.23	...	...	...	...	...	...	3101.0	...	...
	II.....	17.46	4.53	18.73	.686	4.47	−2.72	1	.5	0	3096.6	8.96	*
	III.....	18.27	2.98	19.18	.319	4.03	3.61	1	.436	0	3093.4	6.28	*
	IV.....	Parameters could not be estimated											

6	I.....	31.66	6.73	17.86	...	...	...	...	...	...	3137.4	...	...
	II.....	19.44	4.39	18.23	.236	3.60	4.22	1	.5	0	3131.6	11.58	**
	III.....	19.44	4.31	18.58	.239	3.96	3.80	1	.449	0	3131.2	0.90	...
	IV.....	Parameters could not be estimated											
7	I.....	29.86	5.81	23.04	...	...	...	...	...	...	3096.9	...	...
	II.....	12.83	4.69	20.72	.579	6.15	2.76	1	.5	0	3070.3	53.08	***
	III.....	12.80	4.69	20.72	.420	−6.16	2.75	1	.508	0	3070.2	0.04	...
	IV.....	12.77	4.69	20.72	.415	−6.19	2.69	1	.5	.030	3070.0	0.58	...
8	I.....	28.65	9.30	20.55	...	...	...	...	...	...	3169.8	...	...
	II.....	9.74	9.48	20.78	.603	−5.67	1.95	1	.5	0	3144.9	12.44	**
	III.....	9.60	9.63	20.78	.609	−5.64	2.00	1	.475	0	3144.7	0.30	...
	IV.....	9.84	9.17	20.82	.606	−5.71	1.92	.988	.5	0	3144.8	0.10	...
9	I.....	31.86	4.49	17.18	...	...	...	...	...	...	3093.4	...	...
	II.....	15.30	4.65	20.34	.204	6.15	1.48	1	.5	0	3083.6	19.50	***
	III.....	15.32	4.65	20.34	.796	−6.14	1.48	1	.501	0	3083.6	0	...
	IV.....	14.39	2.03	19.95	.728	−6.63	0.66	.961	.5	.274	3081.1	3.06	...
10	I.....	30.11	5.43	18.50	...	...	...	...	...	...	3091.8	...	...
	II.....	29.05	−2.71	20.65	.456	−1.55	−4.63	1	.5	0	3090.3	3.00	...
	III.....	31.83	3.48	20.73	.435	1.12	−4.28	1	.159	0	3089.9	0.84	...
	IV.....	Parameters could not be estimated											

NOTE: Models: I = classical polygenic; II = Mendelian "mixed" model; III = non-Mendelian, t_2 free; IV = non-Mendelian, t_1 and t_3 free. Significance levels for chi-square: * = .05; ** = .01; *** = .001.

† Indicates parameter fixed on boundary value.

strong claim based on a single large study would have a high probability of replication by the same methods and would be substantively misleading.

It is ill advised to generalize too sweepingly from conclusions based on only 10 simulations but the results add to unease about the value of likelihood statistics computed on the assumption of normality in the residual effects as a guide to the importance of a major gene. The danger is especially great when the methods are applied naïvely to behavioral test scores derived simply by counting the correct responses to a series of binary items. Under these circumstances, the most careful selection of items will generate significant kurtosis, which, even in the absence of significant skewness, may lead mistakenly to the inference that a gene of major effect is contributing to individual differences in behavior.

In the simple case of the ideal test, we have considered it likely that much of the support for a major gene would disappear under an appropriate transformation of the data. Transformations that remove skewness alone, however, are not sufficient to preclude erroneous inferences. Many tests employed for classification and diagnosis, especially in the personality domain, consist of items selected for their ability to discriminate between criterion groups. The item difficulties of such tests will not display a uniform distribution over the entire range of the latent trait but will cluster around the trait value dividing the normal group from the criterion set. It would be no surprise that scores on such a test given to the normal population might lead the geneticist to the conclusion that there is a hidden genetic dichotomy underlying the observed score distribution. It remains to be seen how far mere transformation of the kind employed in analyses of physiological variables would be sufficient to remove support for a major gene.

A general solution to this problem requires the combination of the mixed model with the latent trait model of psychometric theory that estimates abilities from the item parameters and individual response vectors rather than the total scores. Although it is a relatively simple matter to write the model and express the likelihood function, the practical problem of estimation is far from simple because of the large number of parameters involved and the need for repeated numerical integration. As an interim measure, however, we have estimated the subjects' abilities from their responses to the individual items using an algorithm proposed by Birnbaum [7]. As long as the subjects' underlying abilities are normal, it is possible to supply good trial values for subjects' abilities by assigning normal weights to the raw scores and recover estimates from the response vectors. Such estimates do not have the undesirable properties of the raw scores based on summation of correct responses. When the abilities are estimated for the comparatively "good" tests simulated here and for the populations in which the underlying trait is indeed normal, there was no support for a major gene or any nongenetic discontinuity in ability. It is, however, unclear whether this would be generally true for less ideal tests because the precision of ability estimates also depends on the selection of items in the test [7]. Thus, there may be nonnormality in the estimates of ability that would not justify the statistical assumptions implicit in the current formulation of the mixed model.

REFERENCES

1. MORTON NE, MACLEAN, CJ: Analysis of family resemblance. III. Complex segregation of quantitative traits. *Am J Hum Genet* 26:489–503, 1974
2. ELSTON RC: Major locus analysis for quantitative traits. *Am J Hum Genet* 31:655–661, 1979
3. GERRARD JW, RAO DC, MORTON NE: A genetic study of immunoglobulin E. *Am J Hum Genet* 30:46–58, 1978
4. MACLEAN CJ, MORTON NE, LEW R: Analysis of family resemblance. IV. Operational characteristics of segregation analysis. *Am J Hum Genet* 27:365–384, 1975
5. GO RCP, ELSTON RC, KAPLAN EB: Efficiency and robustness of pedigree segregation analysis. *Am J Hum Genet* 30:28–37, 1978
6. LORD FM, NOVICK MR: *Statistical Theories of Mental Test Scores*. Reading, Mass., Addison-Wesley, 1979
7. BIRNBAUM A: Some latent trait models and their use in inferring an examinee's ability, in ibid.
8. MATHER K, JINKS JL: *Biometrical Genetics. The Study of Continuous Variation*. London, Chapman and Hall, 1982
9. NUMERICAL ALGORITHMS GROUP: FORTRAN. Library manual, mark 9, vol 3. Oxford, England, Numerical Algorithms Group, 1982