

THE GENETIC ANALYSIS OF CONTINUOUS VARIATION: A COMPARISON OF EXPERIMENTAL DESIGNS APPLICABLE TO HUMAN DATA

II. ESTIMATION OF HERITABILITY AND COMPARISON OF ENVIRONMENTAL COMPONENTS

By L. J. EAVES

Department of Genetics, University of Birmingham

A method of computer simulation for comparing the relative efficiencies of alternative experimental designs is applied to the study of the efficiency with which broad heritability can be estimated and the two main sources of environmental variation separated in data from human populations. Two minimal sets of data are compared with respect to these criteria and the effects of varying the proportions of the groups of relatives comprising a set are investigated. The relationship between efficiency, the level of broad heritability and the degree of dominance is noted, and the results related to those of a previous study.

1. INTRODUCTION

A previous paper (Eaves, 1969) proposed a method of computer simulation for the comparison of different experimental designs for studies in human psychogenetics. The method was applied to investigate the efficiency with which additive genetic variation could be distinguished from dominance variation in three alternative designs. This criterion was adopted because it is usually the most demanding in any genetic study, but it was noted that other criteria might be adopted to suit the purpose of a particular analysis.

Two other criteria which deserve consideration are the ability of an experiment to detect genetic variation of any kind, and to separate environmental influences acting within from those acting between families. These criteria are the subject of the simulations to be discussed here.

2. METHOD

The biometrical-genetical background to this study and a fuller discussion of the method are given in the previous paper, illustrated with a worked example of the computations. Two minimal sets of data were considered in the present study, each composed of three groups of related pairs of individuals, and each allowing the estimation of the four main components of variation, namely additive (D_R) and non-additive (H_R) genetic variation; variation due to environmental differences acting within families (E_1); and variation due to environmental differences acting between families (E_2).

The two minimal sets of data to be discussed are:

Set 1. Monozygotic twins reared together (MZ_T), monozygotic twins reared apart (MZ_A), and full siblings reared together (S_T).

Set 2. Monozygotic twins reared together, full siblings reared apart (S_A), and full siblings reared together.

The two sets differ, therefore, only in the foster group included in the design, since in Set 2 monozygotic twins reared apart are replaced by full siblings reared apart. This is readily justifiable on economic grounds because MZ_A are extremely difficult to ascertain in large numbers. For each of the three groups of relatives in either set of data, within- and between-pairs analyses of variance can be obtained, providing six second-degree statistics for each set. The expectations of the within- and between-pairs mean squares in terms of D_R , H_R , E_1 and E_2 are given Table 1, on the assumption that mating is at random. The definition of the parameters and the method by which they are combined to provide the expectations of components of variance derivable from the analysis of variance may be found in the work of Mather (1949), and of Jinks & Fulker (1970). The fact that six statistics are predicted by four parameters enables least-squares estimates of the parameters to be obtained, leaving two degrees of freedom for testing the simple genetical and environmental model assumed (Nelder, 1960; Hayman, 1960). The proportion of variation which is

TABLE 1. THE EXPECTATIONS OF THE SECOND-DEGREE STATISTICS OBTAINABLE FROM A WITHIN- AND BETWEEN-PAIRS ANALYSIS OF VARIANCE FOR THE FOUR GROUPS OF RELATIVES CONSIDERED

(The expectation of mean squares are given in terms of the additive genetic component, D_R , the dominance genetic component, H_R , the within-families environmental component, E_1 , and the between-families environmental component, E_2 .)

Mean square	Coefficients of parameters			
	Genetic		Environmental	
	D_R	H_R	E_1	E_2
Monozygotic twins reared together				
Between pairs	1.0	0.5	1.0	2.0
Within pairs	0.0	0.0	1.0	0.0
Monozygotic twins reared apart				
Between pairs	1.0	0.5	1.0	1.0
Within pairs	0.0	0.0	1.0	1.0
Full sibs reared together				
Between pairs	0.75	0.3125	1.0	2.0
Within pairs	0.25	0.1875	1.0	0.0
Full sibs reared apart				
Between pairs	0.75	0.3125	1.0	1.0
Within pairs	0.25	0.1875	1.0	1.0

attributable to both additive and non-additive genetic causes is termed the broad heritability (h_b^2), and may be estimated from

$$\hat{h}_b^2 = (\frac{1}{2}\hat{D}_R + \frac{1}{4}\hat{H}_R) / (\frac{1}{2}\hat{D}_R + \frac{1}{4}\hat{H}_R + \hat{E}_1 + \hat{E}_2),$$

where \hat{D}_R , \hat{H}_R , \hat{E}_1 and \hat{E}_2 are the least-squares estimates of the parameters of the basic model. Provided that the gene frequencies are equal at all the loci involved in the expression of the trait under consideration, the dominance ratio may be estimated from $\sqrt{(\hat{H}_R/\hat{D}_R)}$ (Mather, 1949).

The ability of a given design to detect genetic variation is reflected in the variance of the estimate of broad heritability. The formula for the variance of a ratio given by Kempthorne (1957), and applied by Jinks & Fulker (1970), may be employed to provide an estimate of the variance of the estimate of broad heritability. Where a/b is a ratio, its variance, $V(a/b)$, is approximately given by

$$V(a/b) = (a/b)^2 [V(a)/a^2 - 2W(a . b)/ab + V(b)/b^2],$$

where $V(a)$ is the variance of the numerator of the ratio, $V(b)$ the variance of the denominator, and $W(a . b)$ the covariance between the numerator and denominator.

For the case where a/b is the estimate of broad heritability we have

$$a = \frac{1}{2}\hat{D}_R + \frac{1}{4}\hat{H}_R \quad \text{and} \quad b = \frac{1}{2}\hat{D}_R + \frac{1}{4}\hat{H}_R + \hat{E}_1 + \hat{E}_2,$$

whilst in notation analogous to that of the general formula above,

$$V(a) = \frac{1}{4}V(\hat{D}_R) + \frac{1}{16}V(\hat{H}_R) + \frac{1}{4}W(\hat{D}_R . \hat{H}_R),$$

$$V(b) = \frac{1}{4}V(\hat{D}_R) + \frac{1}{16}V(\hat{H}_R) + V(\hat{E}_1) + V(\hat{E}_2) + \frac{1}{4}W(\hat{D}_R . \hat{H}_R) + W(\hat{D}_R . \hat{E}_1) + W(\hat{D}_R . \hat{E}_2) + \frac{1}{2}W(\hat{H}_R . \hat{E}_1) + \frac{1}{2}W(\hat{H}_R . \hat{E}_2) + 2W(\hat{E}_1 . \hat{E}_2)$$

and

$$W(a . b) = \frac{1}{4}V(\hat{D}_R) + \frac{1}{16}V(\hat{H}_R) + \frac{1}{4}W(\hat{D}_R . \hat{H}_R) + \frac{1}{2}W(\hat{D}_R . \hat{E}_1) + \frac{1}{2}W(\hat{D}_R . \hat{E}_2) + \frac{1}{4}W(\hat{H}_R . \hat{E}_1) + \frac{1}{4}W(\hat{H}_R . \hat{E}_2).$$

The parameters are estimated by the method of weighted least squares and their variances and covariances are the elements of the inverted information matrix derived from the least-squares analysis. Substitution of these values in the above formulae enable the variance of the broad heritability estimate, $V(h_b^2)$, to be calculated. The magnitude of this variance was the first criterion to be investigated by the simulations reported here.

The second criterion to be discussed is the variance of the difference between the estimates of the two environmental components of variation, $V(\hat{E}_1 - \hat{E}_2)$. This is given by the much simpler expression:

$$V(\hat{E}_1 - \hat{E}_2) = V(\hat{E}_1) + V(\hat{E}_2) - 2W(\hat{E}_1 . \hat{E}_2).$$

The terms of this expression are once more the elements of the inverse matrix of the weighted least-squares analysis.

The method described in the previous paper was employed to conduct weighted least-squares analyses on 'data' generated for a hypothetical population of unit total variance. It was assumed for the purposes of this study that the two environmental components accounted for equal proportions of the variance. The broad heritability, h_b^2 , was allowed to take all values from 0.1 to 0.9 in turn by intervals of 0.1. The degree of dominance, given by $\sqrt{(H_R/D_R)}$, was allowed to take successively the three values 0.1, 0.5 and 1.0. These restraints enable values for the four parameters to be calculated. The parameters thus obtained can be combined, according to their coefficients in the expectations of mean squares, to provide the second-degree statistics on which the weighted least-squares analyses can be conducted. For each of the 27 combinations of level of broad heritability and degree of dominance, second-degree statistics were generated for each minimal set of data in turn. These statistics were then treated as if they had been derived by sampling a population and subjected to weighted least-squares analyses to re-estimate the parameters which had been used to generate the 'data'. Superimposed upon the variation of design, heritability and dominance, was the variation in the relative numbers of each of the three groups of which a given design was composed. The total experimental size was assumed to be unity in every case in order to provide overall consistency in the results. The relative proportions of the groups, however, were varied in every possible combination, each group being allowed to be included in every proportion from 0.1 to 0.8, by intervals of 0.1. When the proportions in one group were assigned, those in the other two were allowed to share the remaining part of the sample in varying proportions. Thus in Set 1, for example, when the proportion of MZ_T was fixed at 0.7, the proportions of MZ_A could be either 0.1 or 0.2 with the proportion of the third group (S_T) fixed at 0.2 and 0.1, respectively, by the restraint that the total sample size should be unity.

For each simulated experiment the second-degree statistics were weighted by the reciprocals of their theoretical errors. The information matrix was calculated and this was inverted to give the matrix of the variances and the covariances of the estimates of the parameters. The method of computation is described and illustrated with a worked example in the account of earlier simulation work (Eaves, 1969). It is possible to combine the elements of the variance-covariance matrix to provide the variance of any desired comparison or combination of the estimates obtained from the analysis. In the present case the elements were combined according to the two methods described above to provide the variance of the estimate of broad heritability, $V(\hat{h}_b^2)$, and the variance of the difference between the estimates of the two environmental components, $V(\hat{E}_1 - \hat{E}_2)$. In order to assess the efficiency of the experiment the reciprocals of these two combined variances were taken as the appropriate indices:

$$I_h = 1/V(\hat{h}_b^2) \quad \text{and} \quad I_e = 1/V(\hat{E}_1 - \hat{E}_2).$$

The values of I_h and I_e were computed for both minimal sets of data, for all the conditions of heritability and dominance described above, with all possible

combinations of proportions of the different types of pair within the set. The extensive computations necessary were conducted on the KDF 9 computer at Birmingham University.

3. RESULTS AND DISCUSSION

It will be appreciated that the results of this study are too extensive to present in detail over the whole range investigated. The discussion will be confined to a presentation of the general trends and conclusions, illustrated by a more detailed consideration of some typical examples.

The results of these simulations extend the conclusion of the previous study that the efficiency of an experiment for a given criterion depends upon the relative proportions of the groups of related individuals which form the basis of the sample. Table 2 presents the optimum values of the two indices of efficiency investigated in this study for all the levels of heritability and dominance considered.

TABLE 2. THE OPTIMUM VALUES OBTAINED FOR THE EFFICIENCY OF THE ESTIMATE OF BROAD HERITABILITY, I_h , AND THE EFFICIENCY WITH WHICH E_1 CAN BE SEPARATED FROM E_2 , I_e

(Values are tabulated for two minimal sets of data, for nine levels of broad heritability, h_b^2 , and three levels of dominance, $\sqrt{(H_R/D_R)}$)

		Set 1 (MZ _T , MZ _A , S _T)					
		I_h			I_e		
h_b^2	$\sqrt{(H_R/D_R)} =$	0.1	0.5	1.0	0.1	0.5	1.0
0.1		0.8176	0.8176	0.8176	0.1756	0.1756	0.1756
0.2		0.8736	0.8737	0.8737	0.2084	0.2084	0.2084
0.3		0.9798	0.9799	0.9799	0.2583	0.2583	0.2583
0.4		1.1620	1.1621	1.1623	0.3373	0.3373	0.3373
0.5		1.4765	1.4767	1.4773	0.4706	0.4706	0.4706
0.6		2.0645	2.0591	2.0605	0.7182	0.7183	0.7183
0.7		3.2976	3.2992	3.3026	1.2564	1.2564	1.2565
0.8		6.7444	6.7493	6.7592	2.7979	2.7979	2.7979
0.9		24.7134	24.7385	24.7895	11.1296	11.1296	11.1296
		Set 2 (MZ _T , S _A , S _T)					
0.1		0.1912	0.1907	0.1895	0.1796	0.1790	0.1777
0.2		0.2025	0.2011	0.1982	0.1921	0.1906	0.1876
0.3		0.2194	0.2169	0.2199	0.2106	0.2080	0.2026
0.4		0.2400	0.2357	0.2274	0.2327	0.2283	0.2195
0.5		0.2637	0.2570	0.2440	0.2583	0.2513	0.2379
0.6		0.2906	0.2809	0.2647	0.2860	0.2770	0.2603
0.7		0.3373	0.3235	0.2982	0.3353	0.3211	0.2952
0.8		0.3911	0.3696	0.3317	0.3904	0.3686	0.3302
0.9		0.4457	0.4145	0.3613	0.4456	0.4142	0.3609

The efficiency with which broad heritability is estimated is seen to be a direct function of the level of heritability, since I_h increases as the heritability is increased. In other words the variance of the heritability estimate becomes smaller as the heritability increases. This conclusion is verified by extensive unpublished reanalyses of data from twin studies, in which the variance of an approximate estimate of broad heritability is typically smaller for higher heritabilities. It will be noted that Set 1 is more efficient than Set 2 at all levels of heritability, and that this difference is most marked when the heritability is high. When h^2 is 0.9 the value of I_h obtained from Set 1 is about 60 times as great as that obtained from Set 2. In the situation studied here, where the two environmental components account for equal proportions of the variance, the results obtained for I_e are directly parallel to those obtained for I_h . Set 1 provides more efficient separation of E_1 and E_2 than Set 2.

Variations in the degree of dominance have little effect on the value of I_h . This reflects the fact that the heritability ratio is a general comparison between genetic and environmental variation. The variance of the genetic portion of the variation and its covariance with the environmental portion are largely independent of the relative contributions of additive and non-additive variation to the overall genetic variance. Between dominance at the lowest level, 0.1, and dominance at the highest level, 1.0, the values of I_h obtained vary only in the fourth significant figure for Set 1. The difference is larger for Set 2, but still no more than 25 per cent, even when the heritability is high.

A more detailed examination of the results obtained for Set 1 illustrates an important weakness of this particular method of simulation for the investigation of experimental design. Fig. 1 presents graphically the relationship between the value of I_h and the proportion of MZ_A included in the sample.

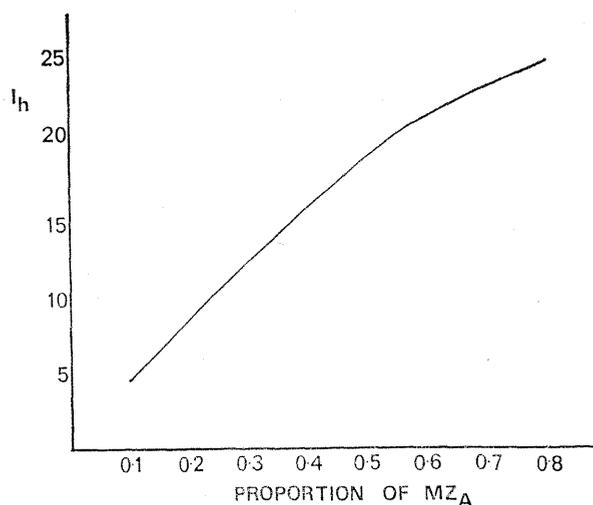


FIGURE 1. A typical graph depicting the variation of the efficiency of estimates of heritability, I_h , for a minimal set of data with varying proportions of monozygotic twins reared apart, MZ_A . The curve is that obtained when the broad heritability is 0.9, and the level of dominance is 0.1.

The efficiency with which the broad heritability can be estimated increases continually with the proportion of MZ_A to the extent that a true optimum has not been reached even when the proportion of MZ_A reached 0.8, at which point the simulation was terminated. If the proportion of MZ_A had been allowed to increase further the efficiency would have also increased, indicating that the most efficient design for the estimation of broad heritability is the limiting one, in which the entire sample is composed of monozygotic twins reared apart. This conclusion is readily explicable in terms of the genetic and environmental influences assumed to be active in the determination of the within- and between-pairs components of variation for monozygotic twins reared apart. The between-pairs component, σ_b^2 , contains all the genetic variation, and σ_w^2 , the within-pairs component, contains all the environmental variation. Thus the intraclass correlation of monozygotic twins reared apart is a direct estimate of the broad heritability. The method of simulation employed automatically confirms that the most efficient design for the estimation of a simple effect is that which enables a direct, perfect fit, solution of the necessary parameters, without regard to any other criteria the experimenter may wish to adopt. Although a direct and efficient estimate of h_b^2 can be obtained from MZ_A alone, such an experiment would allow for neither the distinction between additive and dominant genetic variation, nor for within- and between-family components of environmental variation. A design based solely upon MZ_A will have the further deficiency of precluding a statistical test of the adequacy of the assumptions underlying the simple biometrical model.

A further illustration of the weakness of the method is given by the efficiency with which Set 1 can separate the two environmental components of variation. The most efficient design for the separation of E_1 and E_2 is that which consists of equal proportions of MZ_T and MZ_A . Clearly this is due to the fact that the within-pair variance of the MZ_T is a direct estimate of E_1 , and the within-pair variance of MZ_A is a direct estimate of the total environmental variance, $E_1 + E_2$. Any further information obtained from the inclusion of other groups of relatives is superfluous as far as this criterion is concerned. In this case a test of the biometrical model is still possible, but the two sources of genetic variation would have to be confounded in order for the model to be fitted. No distinction between D_R and H_R could be attempted if no non-twin group were included in the sample. This deficiency in the simulation method means that the results have to be interpreted with proper regard to the purpose of the experiment and the desirability of providing a statistical test of any assumptions made in the definition of the model which is fitted to the data. It would be possible, using a similar simulation technique to the one employed here, to investigate the sensitivity of the different experimental designs to known departures from the assumptions underlying the simple genetical model.

Minimal data Set 2 does not include any group or combination of groups which would allow a more direct solution for either the broad heritability or the two environmental components of variation. The simulation method is therefore

effective in selecting those proportions of the three groups which provide the best estimates of the desired parameters. The graph (Fig. 2) illustrates part of the results obtained when the dominance ratio is assumed to be 0.1 and the broad heritability is set at 0.9. It may be thought of as a section through a surface in three dimensions, of which the two horizontal coordinates represent the proportions of MZ_A and S_A in the sample, and the vertical coordinate the value of I_h . The section has been taken parallel to the ' S_A ' axis, so that it passes through the maximum value of I_h . The striking difference between the form of the first graph (Fig. 1) and that of the second (Fig. 2) is entirely due to the fact that the former represents a situation where a solution is possible with a simpler design than that assumed, whereas the latter does not. The value of I_h has a clear maximum at unique proportions of the three groups of which Set 2 is formed: any deviation from these proportions reduces the efficiency of the experiment.

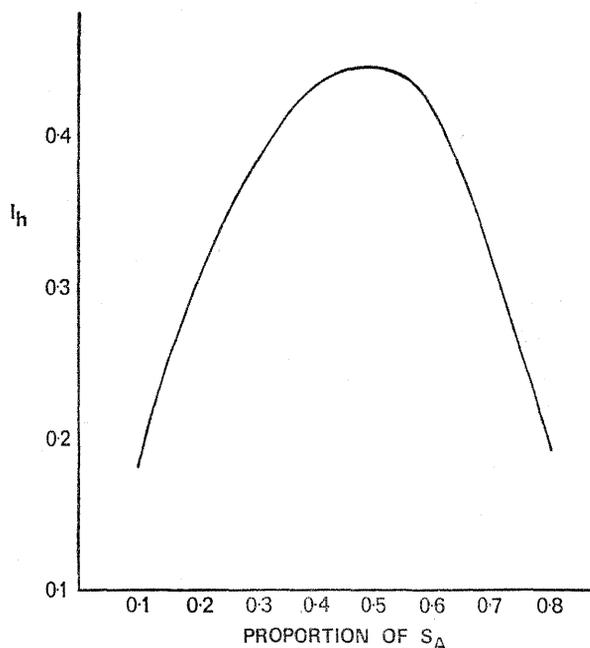


FIGURE 2. A representative graph of the variation of efficiency of heritability estimates for a minimal set of data which includes varying proportions of full siblings reared apart, S_A , in the place of monozygotic twins reared apart. For the purposes of comparison with Fig. 1, the level of broad heritability has been fixed at 0.9, and the level of dominance at 0.1.

It was found for Set 2 that the relative proportions of the three groups required for the optimum estimate of h_b^2 varied with the level of heritability. This conclusion was recognized with regard to the separation of D_R and H_R in the earlier simulation study (Eaves, 1969). It was noted on that occasion that the proportions were sufficiently stable under changing conditions to justify the design of experiments with some regard to the efficiency of the most likely estimates of the parameters required.

The validity of earlier conclusions about the relative merits of foster siblings and separated monozygotic twins must be assessed in the light of the more extensive information made available by the present investigations. It was remarked previously that much effort currently devoted to the accumulation of data on twins might be profitably devoted to sampling full siblings reared apart. No direct comparison of the two minimal sets of data is possible in this case since, as discussed above, Set 1 contains two subsets of data which allow a more direct solution of the two effects under consideration. With respect to the criterion of the efficiency of heritability estimates, it can be seen that the efficiency of estimation increases as the perfect fit situation is approached, at which the model becomes untestable. Further study is needed to elucidate the relationship between the efficiency with which heritability can be estimated, and the sensitivity of the experiment to observed departures from the theoretical expectations on the basis of the assumed biometrical model. It is probable that an inverse relationship exists between efficiency of estimation and testability of the model which demands a compromise solution. If this is assumed, and the proportion of MZ_A is fixed at 0.5 rather than at unity, as the perfect fit solution would require, then the efficiency of Set 1 is still 40 times as great as that of Set 2 when the heritability is 0.9. If economic considerations relating to data collection are excluded from the interpretation of the results, then the weight of the statistical conclusions would favour monozygotic twins reared apart in preference to full siblings reared apart. Since, however, MZ_A pairs are difficult to ascertain, the tolerance of Set 1 to the unavailability of MZ_A is worth indicating. For populations and traits for which the heritability is less than 0.6, and for samples of the same size, the value of I_h for Set 1 when only 10 per cent of the pairs are MZ_A is lower than the corresponding value for Set 2, when the proportion of S_A is maintained at 30 per cent. This suggests that S_A are still an economical alternative to MZ_A , providing that the availability of the former is about three times as great as the availability of the latter. If the heritability is greater than 0.6, the foster-siblings would have to be even more readily available than monozygotic twins reared apart if the substitution is to be justified on both economical and statistical grounds.

If the discussion is restricted to Set 2 alone the most efficient estimates of broad heritability, and the most precise separation of E_1 and E_2 , are obtained when the relative proportions of the three groups are approximately 0.2 MZ_T , 0.5 S_A and 0.3 S_T , for heritabilities in the intermediate range. This result is the same as that obtained earlier for the efficiency with which additive and dominance genetic variation could be separated. For Set 1 to exceed this efficiency a total proportion of 0.6 monozygotic twins (reared together and apart) would be required. This proportion would be prohibitive in any experiment of substantial size.

4. SUMMARY AND CONCLUSIONS

The simulation studies discussed in this paper provide information about the relative efficiency with which two minimal sets of data allow the estimation

of broad heritability and the separation of the two main sources of variation for quantitative traits in human populations.

While it is recognized that a set of data comprising measures made on monozygotic twins reared together, monozygotic twins reared apart and full siblings reared together can provide more efficient estimates of the two effects to be estimated, it is noted that the limited availability of monozygotic twins reared apart may reduce the practical utility of this particular set. It can be shown that the substitution of full siblings reared apart for monozygotic twins reared apart should be considered if the former are more readily available.

A weakness of this simulation method is that it can lead to the selection of an experimental design which, while it is maximally efficient for the criterion under consideration, would preclude the estimation of other important genetic and environmental effects, or a statistical test of the validity of the biometrical model assumed.

ACKNOWLEDGEMENTS

I am indebted to Professor J. L. Jinks for advice and encouragement and to Professor P. L. Broadhurst for comments on the preparation of the paper. This work, which was carried out while the author was in receipt of an S.R.C. Research Studentship, is part of a research programme in psychogenetics supported by the M.R.C.

REFERENCES

- EAVES, J. L. (1969). The genetic analysis of continuous variation: a comparison of experimental designs applicable to human data. I. *Br. J. math. statist. Psychol.* **22**, 131-147.
- JINKS, J. L. & FULKER, D. W. (1970). A comparison of the biometrical genetical, MAVA, and the classical approaches to the analysis of human behaviour. *Psychol. Bull.* **73**, 311-349.
- HAYMAN, B. I. (1960). Maximum likelihood estimation of genetic components of continuous variation. *Biometrics* **16**, 369-381.
- KEMPTHORNE, O. (1957). *An Introduction to Genetic Statistics*. New York: Wiley.
- MATHER, K. (1949). *Biometrical Genetics*. London: Methuen.
- NELDER, J. A. (1960). Estimation of variance components in certain types of experiment in quantitative genetics. In O. Kempthorne (ed.), *Biometrical Genetics*. London: Pergamon Press.