Acta Genet Med Gamellol 30:9-38 (1981)
The Mendel Institute/Alan R. Liss, Inc.

Received 2 September 1980
Final 25 February 1981

A Statistical Model and Analysis for Genetic and Environmental Effects in Responses From Twin-Family Studies

James S. Williams,[1,2] Hariharan Iyer[1]

[1]*Department of Statistics, Colorado State University, Fort Collins, and* [2]*Institute for Behavioral Genetics, University of Colorado, Boulder*

A statistical model and analysis for genetic and environmental effects in twin-family data are presented. The model is used to derive expressions for phenotypic correlations of 22 essential pair relationships in twin-family units. The analysis proceeds in two steps. First, differential effects of sex, generation, and sex-zygosity of twin-family units and correlations due to cluster sampling are eliminated from correlation data. Then, estimates and tests of model parameters are calculated from the adjusted data. The theory and methods were developed for a Swedish twin-family study of many behaviors possibly related to the smoking habit. There, it is important to screen for behaviors that clearly are under genetic control and to assess relative influences of various biological and social environments on the development of all behaviors. Height data from the Swedish study are used to illustrate concepts and methods presented in this paper.

Key words: Twin models, Twin-family studies, Smoking, Genetic screening, Family relationship, Social behavior, Social environment, Height

INTRODUCTION

Two of the broadly stated objectives of the Swedish twin-family study described by Crumpacker et al [5] are to screen many types of behavior that are possibly related to the smoking habit for evidence of genetic control and to assess the relative influences of several biological and social environments on the development of these behaviors. Questionnaire responses by the subjects interviewed in the twin-family sample form much of the data of the study.

There are three fundamental methodological problems to solve before the objectives of any study like the Swedish study can be achieved. The first is to develop sets of questionnaire items that provide quantitative measurements of the required types of behavior. Ideally, the distributional properties of these scored phenotypes are the same for different populations and for all combinations of sex and generation within populations. The prin-

This research was supported by Council for Tobacco Research, U.S.A., Inc. grant CFTR 1066. The study of which it is a part was made possible by collaboration with the Department of Environmental Hygiene of the Karolinska Institute and of the Swedish National Environmental Protection Board.

0001-5660/81/3001-0009$08.00

cipal item sets or scales that were used in the Swedish study have this property. The Comrey and Eysenck Personality Scales which were given [see 4 and 7] have been thoroughly tested by their authors in U.S. and English populations. Translated versions of these have been reported by Floderus [10] and Vandenberg and Price [20] to have similar characteristics in the Swedish urban population from which the twin-family sample was drawn. Williams et al [21] reported that modified versions of smoking-behavior scales devised in London by Russell et al [18] produce the same results in U.S. and Swedish samples as seen in the original English study. Scale measures of several types of alcohol consumption that were used as part of an inventory of Swedish twins [3, 12] have been modified and extended by Pedersen and McClearn [16] and shown to be reproducible in sex by generation subsamples of the twin-family sample. The second problem is to define a statistical model of genetic and environmental effects on a scored phenotype for pedigrees of the type that characterize units of a twin-family sample. The final need is for a statistical test of fit of score data to the model, estimates of model parameters, and tests of hypotheses concerning these. A detailed description of a model and complementary statistical methods for twin-family studies like the Swedish study are the subjects of this paper. These are expansions of the path-diagram and descriptive and verbal presentations given in Crumpacker et al [5].

There are several models currently used in the application of theories of quantitative genetics to the study of human phenotypes. These have been set out in detail in recent exhaustive reviews by Elston and Rao [6] and Boyle and Elston [1]. None has been designed around the stated objectives of the Swedish study, and none has been worked out for very specific pedigrees such as characterize twin-family units. Along with these two guiding features, the model that we propose has the following desirable properties, not found in other works:

1. Every random variable in the representation of a phenotype score is well-defined by a conditional expectation. As a consequence, correlations between any two components of a score can be derived, and no assumptions concerning these need be entertained.

2. Correlations between random variables for two individuals of any pedigree relationship found in twin-family units are derived with a minimum number of clearly stated assumptions and principles. The resulting systems of correlations are non-negative definite and, therefore, internally consistent. They also satisfy known boundary conditions for random mating and perfectly assorting populations, and hence they satisfy external consistency checks.

3. Factors of the environment of the autosomes are divided into two sets to be accounted for and investigated by different methods. First are the factors, such as generation, sex, and type of twin-family unit, of a coarse-grain characterization of the environment, which are of secondary interest and are eliminated by statistical adjustment of phenotype scores. The second are the remaining factors of a fine-grain characterization, which are of primary interest. These, for example, can be associated with descriptions of prenatal and postnatal environments. How the main effects of these and their interactions with genotypes affect correlations between relatives is carefully explained.

4. Up to 18 parameters appear in a set of model expressions for correlations in phenotype scores between relatives in twin-family units. Nine of these pertain to direct effects of the fine-grain environment, and three more are measures of indirect effects brought about by phenotypic assortment and convergence (increase in similarity of spouses) and by a degree of genetic control of family environment. The liberal parameterization of environmental effects, in contrast to the limited number of parameters included to measure genetic

effects, is part of an intentional effort to describe correlations in phenotype scores as fully as possible by common features of the environment before genetic factors are considered. As a consequence, when applied to screening studies, the evidence can be regarded as strong for those phenotypes indicated to be under partial genetic control.

Two sets of methods constitute our statistical analyses. The second of these is a nonlinear least-squares analysis, which is similar in many aspects to one proposed and used by Rao et al [17] in applications of a quantitative genetic model proposed by Morton [13]. The former is a method for obtaining inter-unit correlation estimates of the intra-class correlations described by our model. These estimates are adjusted for differential location and scale effects of the coarse-grain environment and are also partially adjusted for differential correlative effects produced by the environment.

PAIR CHARACTERISTICS IN TWIN-FAMILY UNITS

Twin-family units in the Swedish sample described by Crumpacker et al [5] are formed around a pair of like-sex MZ or DZ twins born between 1911 and 1935. There are, therefore, four basic types of units that can be indexed by sex and zygosity of the twins. The two spouses of the twins and, ideally, all of the adult children of each married pair complete a unit. In fact, some adult children did not participate in that study, but it was required that there be at least one adult child for at least one family in a unit. Most frequently both families were represented by one or more adult children. The exact composition of the sample can be seen in Table 1.

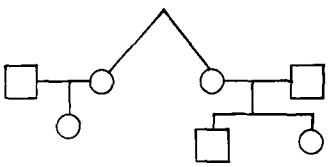
There are 72 different pairs of relatives that can be found in the four types of twin-family units if sexes of the individuals in a pair, as well as sex-zygosity classification of a unit, are considered. Many of the corresponding 72 phenotypic correlations can be combined if only the autosomes are of interest in a genetic analysis and phenotypic scores are adjusted statistically within groups for effects of all combinations of generation, sex, and sex-zygosity of a twin-family unit. The only sex distinctions that remain occur for pairs in which one individual is a twin or twin's spouse and the other is a child. The reason for this is that all such pairs can be traced through the relationship of a child with its mother or with its father and that children can correlate more highly with mothers than with fathers because the two pregnancies of a mother and her child are more closely related than those of a father and his child. In addition to these sex distinctions, a zygosity distinction must be retained for all relationships that can be traced through a pair of twins. Other than these, pair differences can be ignored so that the original 72 pairs can be classified into the following 22 types:

1. MZ twins
2. DZ twins
3. Nontwin siblings
4. Mother and child
5. Father and child
6. Woman and her MZ twin's child
7. Man and his MZ twin's child
8. Woman and her DZ twin's child
9. Man and his DZ twin's child
10. Cousins related through MZ twin mothers

TABLE 1. Relationship Structure of Swedish Sample of Twin-Family Units

Number of children interviewed in pair families[a]	Number of twin-family units related through: MZ twin fathers	MZ twin mothers[b]	DZ twin fathers	DZ twin mothers
0, 1	10	2	7	1
0, 2	4	4	6	0
0, 3	0	0	1	0
1, 1	8	8	6	11
1, 2	9	12	8	12
1, 3	1	4	0	3
1, 4	0	0	1	0
2, 2	5	3	3	3
2, 3	2	2	0	0
3, 3	0	1	0	0
	39	36	32	30

[a]The following is a diagrammatic presentation of a twin-family unit taken from Crumpacker et al [5]:



These families are related through twin mothers. In the first, there is one daughter, and in the second there is a child of each sex. The birth date for the parents is from 1911 to 1935. All children interviewed are at least 20 years old.

[b]One incomplete unit in which one family failed to attend the interview is not counted here. There was one child in the family interviewed. Data from the family were used in our analysis.

11. Cousins related through MZ twin fathers
12. Cousins related through DZ twin mothers
13. Cousins related through DZ twin fathers
14. Husband and wife
15. Individual with MZ twin's spouse
16. Individual with DZ twin's spouse
17. Husband with wife's MZ twin's child
18. Wife with husband's MZ twin's child
19. Husband with wife's DZ twin's child
20. Wife with husband's DZ twin's child
21. Individual with spouse's MZ twin's spouse
22. Individual with spouse's DZ twin's spouse (1)

We next make a further distinction among the 22 listed relationships, which will divide them into two groups. Two individuals are said to be directly related if genes of one are descended from genes of the other. Two individuals are said to be biologically related if they are directly related or if both are directly related to a third individual who was born before either of the two. For example, two children in a family are directly and therefore

biologically related to their parents and biologically related to each other because they are directly related to common parents who were born before them. A husband and wife are directly related to their children, but they are not necessarily biologically related because their children are born after them. Among the 22 relationships, it is clear that the pairs in the first 13 are biologically related. We call these "consanguineous" relationships. It is not possible from the given data to establish a biological relationship for any pair among the final nine relationships. We call these "nonconsanguineous" relationships.

The distinction between consanguineous and nonconsanguineous relationships is used in several steps in the development of our model. Therefore, it is important to notice that in many applications, such as for the Swedish study, the distinction is a feature of the breeding structure of the sampled population as well. In most countries, marriage between first cousins or more closely biologically related people is discouraged and can be legally sanctioned only with a court-approved petition. The most distant biological relationship in the 22 we have listed is between 12 and 13, in which the individuals are first cousins. The closest of the nonconsanguineous relationships is 14. Therefore, in applications we envision, the expected biological relationships, if any, of the nonconsanguineous pair that we will consider is less than the biological relationship of any consanguineous pair in our list.

CORRELATION MODEL

Ours is a second-order statistical model for application to populations at equilibrium. In it, random-variable additive genetic and residual components of phenotype scores are defined subject to specified relationships among the component-score variances and covariances. The covariance relationship between components of a phenotype score is a consequence of the precise definition of the component scores and relates to the definitions of narrow- and broad-sense heritability ratios. The relationships among covariances for the same or different component scores from different individuals partially characterize the mode and effects of population stratification, phenotypic assortment and convergence, and familial genetic influences on residual scores in the population sampled.

Three assumptions that are defined in mathematical terms in subsequent sections on the correlation model are required to develop the relationships of covariances associated with the 22 pairs of individuals listed in (1). They will be described here in less precise terms in order to enhance understanding of their meaning. The first two, which are identical except for score identification, apply to additive genetic and to residual scores of members of nonconsanguineous pairs. They are stated here as variations of a common assumption. The third pertains to the relationship of an individual's residual score to the additive genetic scores of all immediate members of the family in which he was reared.

1. GG(RR). The only direct correlative relationship of additive genetic (residual) scores between nonconsanguineous relatives is between husband and wife. These result from population stratification and phenotypic assortment (and convergence). All other correlative relationships are indirect and can be traced through one or more marriages in a chain of relationships that connect up two nonconsanguineous relatives.
2. GR. The residual score of an individual can be related to the difference in the actual average of his parents' additive genetic scores and the predicted value of the average in terms of the individual's additive genetic score and a similar difference based on the average genetic score of all siblings, other than an MZ twin.

A final assumption used in the development of the model is that correlations in phenotypes are equal for all pairs (among the 72 possible relationships for the four types of twin-family unit) that fit into one of the 22 summary categories of relationships that we have described. For example, the husband-wife correlation does not differ among the four types of units. To approximate the conditions of this assumption, data on phenotype scores must be adjusted for differential effects of generation and sex of individuals and sex-zygosity of family units before any analysis based on the model can be made. These adjustments are discussed in detail in section 4, Statistical Analysis.

Definition of Component Scores

We have elaborated a standard method for the development of operational definitions of random-variable scores [eg, 11: Ch 15] in order to obtain precise mathematical descriptions of causes of phenotypic variation and covariation.

The basic elements of the statistical process that we consider are twin-family units. The variable-length multivariate response for each of these is composed of a specification of the type of unit and, for each person in it, a unique identification tied to a list of phenotype values and characterizations of his or her environment. All variation associated with single responses or pairs of responses is among twin-family units. Random variables that we study pertain to this process and to no other. It would be inappropriate without further study and more information to apply definitions of key components of the model, such as the genetic and environment-by-genotype interaction scores, to other processes.

The first random variables that we introduce are indices of environmental and genetic differences. The first of these is **F**, which is used to make simple, obvious distinctions among possible environments of the autosomes, such as sex, generation, and type of twin-family unit. These are our coarse-grain features of the environment. The second index variable is $\boldsymbol{F}$, which subsumes **F** and is used to introduce additional, finer distinctions in the environment of the autosomes, such as characterizations of prenatal and postnatal familial differences. These are our fine-grain features of the environment. The final variable is **G**, an index of the autosomes' genotype. This is a matrix with zero-one-two elements that indicate the alleles represented at each segregating autosomal locus of the twin-family process.

A suite of score random variables is defined for each value of **F**. The first of the random variables is a phenotype score, which is the deviation of a measurement made on an individual from the conditional expected – that is, the mean value – of the measurement, given **F**. The rest of the random variables are either linear combinations of specific conditional expectations of a phenotype score, best linear predictors of such conditional expectations, or a deviation of a phenotype score from a conditional expected value.

The first component of a phenotype score that we consider is $\mathrm{s}(\boldsymbol{F}, \mathbf{G})$, the conditional expectation of the score given an environment-genotype combination indexed by $\boldsymbol{F}, \mathbf{G}$. Notice that when $\boldsymbol{F}$ is fixed, then **F** also is fixed, because $\boldsymbol{F}$ subsumes **F**. The deviation of a phenotype score from $\mathrm{s}(\boldsymbol{F}, \mathbf{G})$ will be denoted by $\hat{\mathrm{v}}(\boldsymbol{F}, \mathbf{G})$, so that the score itself is the sum $\mathrm{s}(\boldsymbol{F}, \mathbf{G}) + \hat{\mathrm{v}}(\boldsymbol{F}, \mathbf{G})$. The chapeau placed over v is a reminder that, conditionally, the deviation score is a random variable indexed by $\boldsymbol{F}, \mathbf{G}$ in contrast to $\mathrm{s}(\boldsymbol{F}, \mathbf{G})$, which is a mathematical function of the pair. The deviation score is a measure of all effects on a phenotype that are not explained by the genotype and features of the environment indexed by $\boldsymbol{F}$. Among these are measurement errors. The conditional expectation of $\hat{\mathrm{v}}(\boldsymbol{F}, \mathbf{G})$ given $\boldsymbol{F}, \mathbf{G}$ is zero, and therefore the covariance of $\mathrm{s}(\boldsymbol{F}, \mathbf{G})$ with $\hat{\mathrm{v}}(\boldsymbol{F}, \mathbf{G})$ is zero.

The environment-genotype score $s(\boldsymbol{F}, \mathbf{G})$ is used next to define two additional scores. The first of these is a genotype score:

$$s(\mathbf{G}) = E[s(\boldsymbol{F}, \mathbf{G}) \mid \mathbf{F}, \mathbf{G}]. \tag{2}$$

and the second is a fine-grain environment-within-genotype score:

$$s(\boldsymbol{F}, \mathbf{G}) - s(\mathbf{G}). \tag{3}$$

There is no covariance between $s(\mathbf{G})$ and $s(\boldsymbol{F}, \mathbf{G}) - s(\mathbf{G})$ because the conditional expectation of the second of these, given $\mathbf{F}, \mathbf{G}$, for which $s(\mathbf{G})$ is a constant, is zero. Neither of the scores correlates with $\hat{v}(\boldsymbol{F}, \mathbf{G})$ because the conditional expectation of the deviation score given $\boldsymbol{F}, \mathbf{G}$, for which both $s(\mathbf{G})$ and $s(\boldsymbol{F}, \mathbf{G}) - s(\mathbf{G})$ are constants, is zero.

Different types of scores from those just defined can be formed from $s(\mathbf{G})$ and $\mathbf{G}$. These are an additive genetic score $t(\mathbf{G})$, which is the best (minimum mean-squared error) linear predictor of $s(\mathbf{G})$ from the elements of $\mathbf{G}$, and the nonadditive genetic score $u(\mathbf{G})$, which is the difference between $s(\mathbf{G})$ and $t(\mathbf{G})$. The linear predictor is defined with respect to the conditional distribution of $\mathbf{G}$, given $\mathbf{F}$. The covariance between additive and nonadditive genetic scores is zero because a characteristic of best linear predictors is that they do not correlate with their associated errors of prediction. Neither genetic score correlates with $s(\boldsymbol{F}, \mathbf{G}) - s(\mathbf{G})$ or $\hat{v}(\boldsymbol{F}, \mathbf{G})$ for the same reasons that $s(\mathbf{G})$ does not correlate with these scores.

The phenotype score can now be written as a sum of uncorrelated components, viz

$$t(\mathbf{G}) + u(\mathbf{G}) + [s(\boldsymbol{F}, \mathbf{G}) - s(\mathbf{G})] + \hat{v}(\boldsymbol{F}, \mathbf{G}). \tag{4}$$

Our analysis is based on this representation. We combine the final three terms and standardize the resulting two random-variable components by dividing each by its standard deviation. In this fashion, (4) is replaced by the equation

$$X = hY + (1 - h^2)^{1/2}Z \tag{5}$$

where X, Y, and Z are respective standardized phenotype, additive genetic, and residual scores, and h^2, the ratio of variances of additive genetic and phenotype scores, is a narrow-sense heritability coefficient. The residual score Z is a linear combination of uncorrelated nonadditive genetic, fine-grain environment within genotype, and deviation scores. All variables are conditionally defined with respect to a process indexed by $\mathbf{F}$.

If the conditional correlation properties of the processes indexed by $\mathbf{F}$ are identical or are sufficiently similar with respect to their major characteristics, then estimates for correlations that have been averaged over the processes can be fit by modelled expressions for the process correlations. If similarities are not sufficiently strong, then averaged estimates and parameters will not reflect features of environmental and genetic variation in phenotype scores. Therefore, failure of modelled expressions for correlations based on (5) to fit data averaged over coarse-grain categories of the environment is one indication of large interactions of fine-grain environmental and/or genetic effects with coarse-grain environmental effects. Good fit of modelled expressions with parameter estimates that deviate significantly from null values is an indication of negligible interactions with the coarse-grain environment – that is, of a commonality of effects of fine-grain environments and genotypes

among processes indexed by **F**. Good fit with nonsignificant parameter estimates can indicate either the absence of effects or an averaging out of interacting effects.

In place of (5) one might consider

$$X = gY^* + (1 - g^2)^{1/2}Z^* \tag{6}$$

where Y* and Z* are respective standardized genotype and corresponding residual scores and g^2 is a broad-sense heritability coefficient. It turns out that this is not nearly so convenient as the representation in (5). The reason is that Y is a sum of similarly defined additive genetic scores of gametes (Y* is not) and that this property can be used to great advantage in the derivation of correlation scores between relatives in the X, Y, and Z scores.

For any pair of individuals, 1 and 2, there are four correlations that must be described for a second-order statistical model. These are for the pairs (Y_1, Y_2), (Y_1, Z_2), (Z_1, Y_2), and (Z_1, Z_2).

Correlations of Additive Genetic Scores

The index **G** of a diploid genotype of autosomes of an individual can be written as a sum of two indices, $\mathbf{g}_1$ and $\mathbf{g}_2$, for the haploid genotypes of the following:

(1) gametes that fuse to form the individual or
(2) segregation products of any meiotic division in the germ plasm of the individual.

The dimensions of a matrix **g** are the same as those of **G**, and its elements are zero-one random variables, which indicate those alleles at segregating loci that are present. Consequences of the two facts that $\mathbf{G} = \mathbf{g}_1 + \mathbf{g}_2$ and that $t(\cdot)$ is a linear function in the elements of its matrix argument and of population equilibrium are these:

(3) $t(\mathbf{G}) = t(\mathbf{g}_1) + t(\mathbf{g}_2)$ and
(4) because of the mendelian laws of segregation, the correlation between $t(\mathbf{g}_1)$ and $t(\mathbf{g}_2)$ is the same for both identifications (1 and 2, above).

These results hold only for additive genetic scores and make these much easier to work with than either nonadditive genetic or genotype scores. Because of them, one can express the additive genetic score of an individual, Y, as a linear combination of the additive genetic score, Y′, of one of his parents and an uncorrelated residual Z′, viz,

$$Y = \tfrac{1}{2}Y' + \tfrac{1}{2}[(1 - m)/(1 + m)]^{1/2}Z'. \tag{7}$$

The correlation coefficient, m, is Wright's inbreeding coefficient for additive genetic scores. Furthermore, Y′ and Z′ can be written as linear combinations of standardized additive genetic scores, y_1' and y_2', of the gamete that enters into the union that produces the individual and its segregation complements, viz,

$$Y' = \left[\frac{1}{2(1 + m)}\right]^{1/2}(y_1' + y_2') \text{ and } Z' = \left[\frac{1}{2(1 - m)}\right]^{1/2}(y_1' - y_2'). \tag{8}$$

The important feature of this representation is that Z′ does not correlate with any additive genetic score other than that of the individual. The reason for this is that y_1 and y_2 correlate equally with any additive genetic score other than Y, and therefore any correlation

with $y_1 - y_2$ is zero. When convenient in a correlation analysis, Y can be written in the form of (7), and the second term can be ignored.

The correlation coefficients for the 22 pair relationships we must consider can be written as simple functions of four correlation parameters: $\rho, \rho', \rho'', \rho'''$. The derivation is easy if the following order is used: the score of a child is always replaced by the first term in (7), and whenever a new unknown correlation is required, one of the four parameters is introduced. The following completely general results for pairs [on left, refer to (1)] and correlations (on right) are obtained:

1	1
14, 15	ρ
2, 3, 4, 5, 6, 7	$(1 + \rho)/2$
16	ρ'
8, 9	$(1 + \rho + 2\rho')/4$
17, 18	$(\rho + \rho')/2$
21	ρ''
10, 11	$(1 + 2\rho + \rho'')/4$
22	ρ'''
12, 13	$(1 + \rho + 4\rho' + 2\rho''')/8$
19, 20	$(\rho' + \rho''')/2$

(9)

It can be seen that the four parameters are correlations in additive genetic scores between nonconsanguineous relatives in the parent generation. In populations at equilibrium, the husband-wife correlation, ρ, is a function of m – viz, $\rho = 2m/(1 + m)$; in general, the others have no such simple expression.

At this point, we introduce assumption GG, previously described, in order to reduce the number of correlation parameters to one. Without such parsimony, the finished model would contain too many parameters to obtain a unique fit to experimental data. In our precise formulation of GG, Y_1 and Y_2 are additive genetic scores of spouses, and $\mathbf{Y}_1$ is the set of additive genetic scores of all nonconsanguineous relatives of spouse 1. The assumption is that

$$E(Y_1 \mid \mathbf{Y}_1) = \rho Y_2, \tag{10}$$

that is, the conditional expected value of an individual's additive genetic score, given such scores for all his nonconsanguineous relatives, is proportional to the additive genetic score of his spouse. An immediate consequence of (10) is an expression for the correlation coefficient for a husband-wife pair, viz:

$$E(Y_1 Y_2) = E[E(Y_1 Y_2 \mid \mathbf{Y}_1)] = E[Y_2 E(Y_1 \mid \mathbf{Y}_1)] = E(\rho Y_2^2) = \rho. \tag{11}$$

This is the required result for husband-wife and an individual with MZ twin's spouse pairs [pairs 14 and 15 of (1)]. Almost as simple a derivation can be obtained for an individual with DZ twin's spouse [pair 16 of (1)] if we index the twin's spouse, DZ twin, and individual by 1, 2, and 3, respectively:

$$E(Y_1 Y_3) = E[E(Y_1 Y_3 \mid \mathbf{Y}_1)] = E[Y_3 E(Y_1 \mid \mathbf{Y}_1)] = E(\rho Y_3 Y_2) = \rho[(1 + \rho)/2]. \tag{12}$$

https://doi.org/10.1017/S0001566000006590 Published online by Cambridge University Press

This is the expression for ρ' in (9) in terms of ρ. Notice that $E(Y_3Y_2)$ in this derivation is the correlation for DZ twins. The remaining two correlation equations that define parameters can be derived as easily as these. Each turns out to be just the product of coefficients along a direct chain of pair relationships linking up the two individuals; viz:

$$\rho'' = \rho^2 \text{ and } \rho''' = \rho^2[(1 + \rho)/2]. \tag{13}$$

The 22 required correlation coefficients of additive genetic scores, expressed as functions of ρ alone, are displayed later, in Table 3 under the column headed "c_1: Coefficient of h^2." It is easy to show that this system of correlations is positive-definite for any value of ρ between -1 and 1 and therefore is mathematically consistent. They also satisfy the known boundary conditions for no assortment of additive genetic scores ($\rho = 0$) and for perfect assortment of scores ($\rho = -1$ and 1).

Correlations of Additive Genetic With Residual Scores

There are two reasons why additive genetic scores can correlate with residual scores in the 22 pair relationships we consider. The first is that the constructive method used to define additive and nonadditive genetic scores does not necessarily result in their being uncorrelated, except in the pairings of an individual with himself and of MZ twins, which genetically is an equivalent. The second is that the environment of an individual, which influences development of his residual score, can be in part controlled by his immediate relatives and therefore influenced by their genotypes. In the second case, if the influence of genes on environment is measured by the nonadditive score, then this effect will be picked up in the correlations between residual scores. If it is measured by the additive score, then it must appear in a cross correlation of additive genetic with residual scores.

Expressions for the cross correlations can be derived from a precise statement of the previously described assumption GR. We formulate this in terms of the residual score, Z, of an individual and the set, **Y**, of additive genetic scores of all of his consanguineous relatives. The assumption is that

$$E(Z \mid \mathbf{Y}) = \kappa_1[\bar{Y}' - [(1 + \rho)/2]Y] \tag{14}$$

if the individual is an only child or an MZ twin with only one sibling, or

$$E(Z \mid \mathbf{Y}) = \kappa_1[\bar{Y}' - [(1 + \rho)/2]Y] + \kappa_2[\bar{Y} - [(1 + \rho)/2]Y] \tag{15}$$

if the individual is not an MZ twin and has one or more siblings or is an MZ twin with two or more siblings. In these equations, κ_1 and κ_2 are constants, $\bar{Y}'$ is the average of the individual's parents' additive genetic scores, and $\bar{Y}$ is the average of additive genetic scores for all siblings of the individual, other than an MZ twin. Equations (14) and (15) can be combined into a single equation; viz:

$$E(Z \mid Y, \hat{\eta}) = \kappa_1[\bar{Y}' - [(1 + \rho)/2]Y] + \hat{\eta}\kappa_2[\bar{Y} - [(1 + \rho)/2]Y] \tag{16}$$

if we let $\hat{\eta}$ be an independent random-variable indicator of which of equations (14) ($\hat{\eta} = 0$) and (15) ($\hat{\eta} = 1$) is required for a particular problem.

For an individual who is an only child or an MZ twin in a family of two children, the conditional expectation (16) depends only on the difference between the average of

https://doi.org/10.1017/S0001566000006590 Published online by Cambridge University Press

parents' additive genetic scores and the best linear predictor of that score in the individual's additive genetic score. Otherwise, the difference between the average of scores for siblings other than an MZ twin and the best linear predictor also enters into the formula for the conditional expectation. The reason that the additive genetic score of an MZ twin is excluded from these relationships is that it does not differ from Y, and therefore it should have no correlative effect on Z, because Y and Z do not correlate.

We recall that, as a consequence of the manner in which they are constructed, the additive genetic and residual scores in the partition of a phenotype score do not correlate. The GR assumption leads to the same result because

$$\begin{aligned} E(YZ \mid \hat{\eta}) &= E[E(YZ \mid \mathbf{Y}, \hat{\eta})] = E[YE(Z \mid \mathbf{Y}, \hat{\eta})] \\ &= \kappa_1\left\{E(Y\overline{Y}') - [(1+\rho)/2]E(Y^2)\right\} + \hat{\eta}\kappa_2\left\{E(Y\overline{Y}) - [(1+\rho)/2]E(Y^2)\right\} \\ &= \kappa_1\left\{[(1+\rho)/2] - [(1+\rho)/2]\right\} + \hat{\eta}\kappa_2\left\{[(1+\rho)/2] - [(1+\rho)/2]\right\} = 0. \end{aligned} \tag{17}$$

Equation (16) indicates that a parent's additive genetic score can correlate with any of his or her children's residual scores. This does occur because

$$\begin{aligned} E(Y'Z \mid \hat{\eta}) &= E[E(Y'Z \mid \mathbf{Y}, \hat{\eta})] = E[Y'E(Z \mid \mathbf{Y}, \hat{\eta})] \\ &= \kappa_1\left\{E(Y'\overline{Y}') - [(1+\rho)/2]E(Y'Y)\right\} + \hat{\eta}\kappa_2\left\{E(Y'\overline{Y}) - [(1+\rho)/2]E(Y'Y)\right\} \\ &= \kappa_1\left\{[(1+\rho)/2] - [(1+\rho)/2]^2\right\} + \hat{\eta}\kappa_2\left\{[(1+\rho)/2] - [(1+\rho)/2]^2\right\} \\ &= (\kappa_1 + \hat{\eta}\kappa_2)[(1-\rho^2)/4]. \end{aligned} \tag{18}$$

If no distinction is made about the number of children in a family or their twin status, then the expectation of $E(Y'Z \mid \hat{\eta})$ with respect to $\hat{\eta}$ can be used in a model for cross correlations. This we denote by

$$E(Y'Z) = (\kappa_1 + \eta\kappa_2)[(1-\rho^2)/4] = \delta. \tag{19}$$

It is a parameter in our expression for the cross correlations. The complementary correlation of a child's additive genetic score with his parent's residual score is zero, as is indicated by the absence of a child's score in equation (16); ie,

$$\begin{aligned} E(YZ' \mid \hat{\eta}) &= E[E(YZ' \mid \mathbf{Y}', \hat{\eta})] = E[YE(Z' \mid \mathbf{Y}', \hat{\eta})] \\ &= \kappa_1\left\{E(Y\overline{Y}'') - [(1+\rho)/2]E(YY')\right\} + \hat{\eta}\kappa_2\left\{E(Y\overline{Y}') - [(1+\rho)/2]E(YY')\right\} \\ &= \kappa_1\left\{[(1+\rho)/2]^2 - [(1+\rho)/2]^2\right\} + \hat{\eta}\kappa_2\left\{[(1+\rho)/2]^2 - [(1+\rho)/2]^2\right\} \\ &= 0. \end{aligned} \tag{20}$$

In this, $\overline{Y}''$ is an average of grandparents' additive genetic scores, and $\overline{Y}'$ is an average of scores for parent's siblings, other than an MZ twin.

The other pair relationship that defines a parameter of the cross-correlation expressions is the full-sibling pair. For this,

$$\begin{aligned} E(Y_1Z_2) &= E[E(Y_1Z_2 \mid Y_2)] = E[Y_1E(Z_2 \mid Y_2)] \\ &= \kappa_1\left\{E(Y_1\overline{Y}_2') - [(1+\rho)/2]E(Y_1Y_2)\right\} + \kappa_2\left\{E(Y_1\overline{Y}_2) - [(1+\rho)/2]E(Y_1Y_2)\right\} \\ &= \kappa_1\left\{[(1+\rho)/2] - [(1+\rho)/2]^2\right\} + \kappa_2\left\{E[(1+\rho)/2] + [(1-\rho)/2(k-1)] - [(1+\rho)/2]^2\right\} \\ &= (\kappa_1 + \kappa_2)[(1-\rho^2)/4] + \kappa_2E[1/(k-1)][(1-\rho)/2] \\ &= \delta + \kappa_2\left\{(1-\eta)[(1+\rho)/2] + E[1/(k-1)]\right\}[(1-\rho)/2] \end{aligned} \tag{21}$$

https://doi.org/10.1017/S0001566000006590 Published online by Cambridge University Press

where $k - 1$ is the random variable number of siblings of a child, other than an MZ twin, and $k - 1 \geqslant 1$. The second term in the final expression on the right-hand side of (21) is the second parameter in our expressions. We denote it by Δ.

The full set of cross correlations for the 22 pair relationships can be derived in the manner that we have used for (17)–(21). For each pair involving a parent and a child, the complementary correlations must be worked out as in (19) and (20). In the list that follows, these are given with the combination of a parent's additive genetic score with a child's residual score in the first position.

1, 14, 15, 16, 21, 22	0
2, 3	$\delta + \Delta$
10, 11	$(1 + \rho)\delta/2$
12, 13	$(1 + \rho)^2\delta/4$
4, 5, 6, 7	$\delta, 0$
8, 9	$(1 + \rho)\delta/2, (1 + \rho)(\delta + \Delta)/2$
17, 18	$\rho\delta, 0$
19, 20	$\rho(1 + \rho)\delta/2, 0$

(22)

Sums of these terms appear later, in Table 3 under the column headed "c_2: Coefficient of $h(1 - h^2)^{1/2}$." For pair relationships involving two parents or two children, the entry is twice the single value given in the list in (22). For other pairs, the entry is the sum of two values shown in (22).

The results for cross correlations do satisfy known boundary conditions. All of the correlations are zero if neither parents nor siblings have a direct effect ($\kappa_1 = \kappa_2 = 0$) or if additive genetic assortment is positive and perfect ($\rho = 1$). If siblings have no direct effect ($\kappa_2 = 0$) or additive genetic assortment is positive and perfect ($\rho = 1$), then the sum of full sibling correlations is twice the sum of parent with his or her child correlations ($\Delta = 0$).

Correlations of Residual Scores

There is no basis, such as the mendelian law of segregation, on which to construct an analysis of correlations of pairs of residual scores. Therefore, any model to describe these correlations adequately must contain many more parameters than are required for pairs of additive genetic scores. We divide the problem here into two parts. The first concerns correlations for the 13 consanguineous pair relationships in (1), and the second concerns those for the remaining nonconsanguineous pairs.

One feature of the set of consanguineous pairs that can be identified and used is that pairs can be distinguished by relationships between members of a pair in prenatal environments, postnatal familial environments, and nonadditive genetic scores. The first two of these are fine-grain distinctions in the autosomes' environment. For example, effects on adult phenotype of prenatal and postnatal familial environments for MZ and for DZ twins are quite similar and will be treated as the same. However, there are correlation differences in nonadditive genetic scores, because the scores of MZ twins are identical and those of DZ twins are not. In general, differences can be indexed, say by variables f_1, f_2, ℓ, respectively, for prenatal environments, postnatal familial environments, and genotypes. The f variables are elements of $\boldsymbol{F}$. Usually, a correlation in residual scores can be expressed exactly as a Fourier expansion of the following form:

$$\sum_i \sum_j \sum_k \mu_{ijk} \nu_i(f_1) \nu_j(f_2) \nu_k(\ell). \tag{23}$$

https://doi.org/10.1017/S0001566000006590 Published online by Cambridge University Press

In this, a $\nu(\cdot)$ is a basis function for the expansion and a μ is a parameter determined by the expansion. The number of terms required in an expansion depends on the range of the f and ℓ indices.

If there is an expansion of the form in (23) in which all terms except a few of the leading ones are negligible, then the function can be well approximated by

$$\alpha_i + \beta_j + \gamma_k \tag{24}$$

where i, j, and k replace the indices f_1, f_2, and ℓ; and α, β, γ are parameters of the representation. For the 13 consanguineous pair relationships, there are four distinguishable types of association of prenatal environments, five of postnatal familial environments, and four of genotypes. Therefore, we take i = 0, 1, 2, 3; j = 0, 1, 2, 3, 4; and k = 0, 1, 2, 3, and use increasing values of an index to indicate increasing degrees of relationship. For example, for the consanguineous pairs in which individuals are least related – viz, first cousins related through DZ twin fathers [13 of (1)], we have, i = j = k = 0, and the parameterized form of the correlation in residual scores is $\alpha_0 + \beta_0 + \gamma_0$. The next most distantly related pairs are first cousins related through DZ twin mothers [12 of (1)]. The prenatal environments of these cousins are more closely related than those whose fathers are DZ twins, because the mothers who provide the environment are consanguineous relatives. Therefore, we replace α_0 by α_1. The relationship of postnatal familial environments does not depend on the sex of DZ twin parents, nor does the relationship of nonadditive genetic scores. Therefore, there is no change in β and γ parameters, and the parametric expression for the correlation in residual scores is $\alpha_1 + \beta_0 + \gamma_0$. The indices must attain their maximum values for the MZ twins pairs [1 in (1)] in which prenatal environments are identical; postnatal, familial environments are most similar; and nonadditive genetic scores are identical. Therefore, the parametric expression for these pairs is $\alpha_3 + \beta_4 + \gamma_3$.

The exact definitions of the levels i, j, k, which we use for Swedish-type studies, are set out in Table 2. The parametric expressions for the correlations in residual scores for consanguineous relatives are listed under the column headed "c_3: Coefficient of $(1 - h^2)$" in Table 3.

Fine-grain environment by genotype interactions are a cause of errors in the approximation of (23) by (24). These can have two effects on analyses based on this model. First, if substantial interactive effects are unconfounded with additive effects that are accounted for by the approximation, then data will tend not to fit the model. Second, confounded interactive effects are attributed to additive effects and will bias upward estimates of fractions of variance due to the latter. An adequate test of fit is a protection against the first of these. The second is recognized in our applications of the model where we acknowledge that a nonzero variance fraction associated with the fine-grain environment means that the effect is real for at least one genotype, although it could have a zero expectation over all genotypes.

At this point we introduce previously defined assumption RR to complete the derivation of correlations in residual scores for nonconsanguineous relatives. In our precise formulation of RR, Z_1 and Z_2 are residual scores of spouses, and $\mathbf{Z}_1$ is the set of residual scores of all nonconsanguineous relatives of spouse 1. The assumption is that

$$E(Z_1 \mid \mathbf{Z}_1) = \theta Z_2, \tag{25}$$

https://doi.org/10.1017/S0001566000006590 Published online by Cambridge University Press

*TABLE 2. Indices for Fine-Grain Environment of the Swedish Twin-Family Study and Nonadditive Genetic Scores**

Factor (model parameter)	Index value	Degree of similarity
Prenatal familial environment (α)	0	Less similar than for pregnancies of sibling mothers
	1	As similar as pregnancies of sibling mothers
	2	As similar as two pregnancies of one mother or of different pregnancies of MZ twin mothers
	3	As similar as the common pregnancy of twin siblings
Postnatal familial environment (β)	0	Less similar than for double first cousins reared separately in natural homes
	1	As similar as double first cousins reared separately in natural homes
	2	As similar as parent and offspring reared separately in natural home
	3	As similar as nontwin full siblings reared together in natural home
	4	As similar as twins reared together in natural home
Nonadditive genetic score (γ)	0	Biologically related as first cousins
	1	Biologically related as half-siblings
	2	Biologically related as full siblings
	3	Biologically related as MZ twins

*Index values denote different degrees of similarity in environments and nonadditive genetic scores. Only differences in pre- and postnatal environments are considered. These are associated with comparisons among twin-family units, between families in units, and among children within families. In each case, comparisons are made within a type of twin-family unit characterized by the sex and zygosity of twin parents.

that is, the conditional expected value of an individual's residual score, given such scores of all of his nonconsanguineous relatives, is proportional to the residual score of his spouse. The proportionality constant is the correlation for husband-wife pairs; ie,

$$E(Z_1Z_2) = E[E(Z_1Z_2 \mid \mathbf{Z}_1)] = E[Z_2E(Z_1 \mid \mathbf{Z}_1)] = E(\theta Z_2^2) = \theta. \tag{26}$$

The remaining expressions are as easy to derive as this one. Each turns out to be a product of coefficients in a direct chain of pair relationships between the two individuals. The complete list is set out in Table 3 in the final nine entries under the heading "c_3: Coefficient of $(1 - h^2)$."

In all, nine parameters are required for applications of our model of correlations in residual scores to the twin family units of Swedish-type samples. These are the linear combinations: $(\alpha_0 + \beta_0 + \gamma_0)$, $(\alpha_1 - \alpha_0)$, $(\alpha_2 - \alpha_1)$, $[(\beta_1 - \beta_0) + (\gamma_1 - \gamma_0)]$, $[(\beta_2 - \beta_1) + (\gamma_2 - \gamma_1)]$, $(\beta_3 - \beta_2)$, $(\gamma_3 - \gamma_2)$, $[(\alpha_3 - \alpha_2) + (\beta_4 - \beta_3)]$, and θ. (The 13 basic parameters cannot be cleanly separated in the 22 pair relationships we consider.) This is a sharp contrast to the single parameter needed to model correlations in the additive genetic scores. It is due principally to the lack of a biological system with properties that make it possible to relate residual scores of consanguineous relatives and to our desire to construct a model in which correlations in phenotype scores can be closely approximated by correlations in residual scores alone.

*TABLE 3. Model Expressions for Correlations Between Additive Genetic Scores (c_1), the Sum of Complementary Correlations Between Additive Genetic and Residual Scores (c_2), and Between Residual Scores (c_3)**

Pair relationship	c_1: Coefficient of h^2	c_2: Coefficient of $h(1-h^2)^{1/2}$	c_3: Coefficient of $(1-h^2)$
Consanguineous			
1	1	0	$(\alpha_3+\beta_4+\gamma_3)$
2	$(1+\rho)/2$	$2(\delta+\Delta)$	$(\alpha_3+\beta_4+\gamma_2)$
3	$(1+\rho)/2$	$2(\delta+\Delta)$	$(\alpha_2+\beta_3+\gamma_2)$
4	$(1+\rho)/2$	δ	$(\alpha_1+\beta_2+\gamma_2)$
5	$(1+\rho)/2$	δ	$(\alpha_0+\beta_2+\gamma_2)$
6	$(1+\rho)/2$	δ	$(\alpha_1+\beta_2+\gamma_2)$
7	$(1+\rho)/2$	δ	$(\alpha_0+\beta_2+\gamma_2)$
8	$(1+\rho)^2/4$	$(1+\rho)(\delta+\Delta/2)$	$(\alpha_1+\beta_1+\gamma_1)$
9	$(1+\rho)^2/4$	$(1+\rho)(\delta+\Delta/2)$	$(\alpha_0+\beta_1+\gamma_1)$
10	$(1+\rho)^2/4$	$(1+\rho)\,\delta$	$(\alpha_2+\beta_1+\gamma_1)$
11	$(1+\rho)^2/4$	$(1+\rho)\,\delta$	$(\alpha_0+\beta_1+\gamma_1)$
12	$(1+\rho)^3/8$	$(1+\rho)^2\delta/2$	$(\alpha_1+\beta_0+\gamma_0)$
13	$(1+\rho)^3/8$	$(1+\rho)^2\delta/2$	$(\alpha_0+\beta_0+\gamma_0)$
Nonconsanguineous			
14	ρ	0	θ
15	ρ	0	$\theta(\alpha_3+\beta_4+\gamma_3)$
16	$\rho(1+\rho)/2$	0	$\theta(\alpha_3+\beta_4+\gamma_2)$
17	$\rho(1+\rho)/2$	$\rho\delta$	$\theta(\alpha_1+\beta_2+\gamma_2)$
18	$\rho(1+\rho)/2$	$\rho\delta$	$\theta(\alpha_0+\beta_2+\gamma_2)$
19	$\rho(1+\rho)^2/4$	$\rho(1+\rho)\delta/2$	$\theta(\alpha_1+\beta_1+\gamma_1)$
20	$\rho(1+\rho)^2/4$	$\rho(1+\rho)\delta/2$	$\theta(\alpha_0+\beta_1+\gamma_1)$
21	ρ^2	0	$\theta^2(\alpha_3+\beta_4+\gamma_3)$
22	$\rho^2(1+\rho)/2$	0	$\theta^2(\alpha_3+\beta_4+\gamma_2)$

*The index of pair relationships is given in the list in (1). The model expression for a correlation between phenotype scores for any pair is the sum $h^2c_1 + h(1-h^2)^{1/2}c_2 + (1-h^2)c_3$, with the values of c_1, c_2, and c_3 taken from the row indexed by the pair relationship.

It is easy to show that the system of 22 correlations in residual scores is positive-definite for any θ from -1 to 1 if the subsystem for 13 consanguineous pair relationships is positive-definite. This also is required for a mathematically consistent model. The subsystem will always be non-negative-definite if the parameters satisfy the inequality relationships:

$$0 \leqslant \alpha_0 \leqslant \alpha_1 \leqslant \alpha_2 \leqslant \alpha_3, 0 \leqslant \beta_0 \leqslant \beta_1 \leqslant \beta_2 \leqslant \beta_3 \leqslant \beta_4, \text{ and } 0 \leqslant \gamma_0 \leqslant \gamma_1 \leqslant \gamma_2 \leqslant \gamma_3. \quad (27)$$

These are equivalent to a statement that all correlative relationships in effects of environment and nonadditive genetic scores are non-negative. Negative effects, which might occur, for example, in complementary behavior, are not excluded, but there is no property simpler than non-negative definiteness of the subsystem to place bounds on how they are reflected in the correlation parameters other than θ.

The derived expressions for nonconsanguineous pairs of relatives satisfy required boundary conditions if there is no assortment for residual scores ($\theta = 0$) or if assortment is perfect ($\theta = -1$ or 1).

https://doi.org/10.1017/S0001566000006590 Published online by Cambridge University Press

Correlations of Phenotype Scores

The correlation coefficient for phenotype scores X_1 and X_2 of any two individuals in a pair relationship is

$$E(X_1X_2) = h^2E(Y_1Y_2) + h(1 - h^2)^{1/2}[E(Y_1Z_2) + E(Z_1Y_2)] + (1 - h^2)E(Z_1Z_2). \tag{28}$$

Expression for this can be obtained for pairs in list (1) from results summarized in Table 3. For example, the respective correlations for MZ twins and DZ twins [1 and 2 in (1) and Table 3] are

$$h^2 + (1 - h^2)(\alpha_3 + \beta_4 + \gamma_3) \tag{29}$$

and

$$h^2[(1 + \rho)/2] + 2h(1 - h^2)^{1/2}(\delta + \Delta) + (1 - h^2)(\alpha_3 + \beta_4 + \gamma_2). \tag{30}$$

The full system of 22 correlations of this type must be non-negative-definite for an internally consistent model. We have already stated that this property obtains for each of the systems in $E(Y_1Y_2)$ and $E(Z_1Z_2)$ expressions, and therefore it also holds for $E(X_1X_2)$ if $\delta = \Delta = 0$. More generally, one can show that the combined system of $E(Y_1Y_2)$, $E(Y_1Z_2)$, $E(Z_1Y_2)$, $E(Z_1Z_2)$ expression can be non-negative-definite for nonzero δ and Δ.

Useful functions of the model correlation expressions can be derived and interpreted from considerations of equations (4), (5), and (6) and the definitions of indexed α, β, and γ parameters. First, the variance of a nonstandardized phenotype score can be written as a sum of variances as a consequence of correlation properties of the component terms in (4). It is

$$v = \text{var}[t(\mathbf{G})] + \text{var}[u(\mathbf{G})] + \text{var}[s(\boldsymbol{F}, \mathbf{G}) - s(\mathbf{G})] + \text{var}[\hat{v}(\boldsymbol{F}, \mathbf{G})]. \tag{31}$$

The first three terms in (4) are the same for two MZ twins, and the final random variables do not correlate if indices of all environments that create correlations between MZ twins are contained in $\boldsymbol{F}$. In these cases the covariance of phenotype scores for MZ twins is

$$c = \text{var}[t(\mathbf{G})] + \text{var}[u(\mathbf{G})] + \text{var}[s(\boldsymbol{F}, \mathbf{G}) - s(\mathbf{G})] \tag{32}$$

and the correlation is the ratio c/v. The first functions we consider are the fractions of phenotype variance attributed to variation in additive genetic, nonadditive genetic, and genotype scores. The ratio for additive genetic scores, which also appears in (5) and is defined there, is

$$h^2 = \text{var}[t(\mathbf{G})]/v. \tag{33}$$

The ratio for nonadditive genetic scores, which was introduced in the definition of γ_3, is

$$(1 - h^2)\gamma_3 = \text{var}[u(\mathbf{G})]/v. \tag{34}$$

https://doi.org/10.1017/S0001566000006590 Published online by Cambridge University Press

With these two and the definition introduced with equation (6) for the ratio for genotype scores, we have

$$g^2 = \text{var}[s(\mathbf{G})]/v = \left\{\text{var}[t(\mathbf{G})] + \text{var}[u(\mathbf{G})]\right\}/v = h^2 + (1 - h^2)\gamma_3. \tag{35}$$

This leaves

$$(1 - h^2)(\alpha_3 + \beta_4) = \text{var}[s(\boldsymbol{F}, \mathbf{G}) - s(\mathbf{G})]/v \tag{36}$$

for the fraction that can be attributed to all effects of the fine-grain environment. Here, it is important to observe two things. First, $s(\boldsymbol{F}) - s(\mathbf{F})$ rather than $s(\boldsymbol{F}, \mathbf{G}) - s(\mathbf{G})$ is the commonly employed definition of a score for environmental effects, and $s(\boldsymbol{F}) - s(\mathbf{F}) = 0$ implies that $\text{var}[s(\boldsymbol{F}, \mathbf{G}) - s(\mathbf{G})] = 0$ only if there is no fine-grain environment by genotype interaction. Therefore, nonzero $\alpha_3 + \beta_4$ indicates only that there are effects of the fine-grain environment for some genotypes. Second, the notation $\alpha_3 + \beta_4$ suggests that there is no interaction of effects of different types of the fine-grain environment. We have already stated that this is an assumption and indicated consequences of incorrect applications of the assumption. A final variance fraction is

$$1 - (c/v) = (v - c)/v = \text{var}[\tilde{v}(\boldsymbol{F}, \mathbf{G})]/v = (1 - h^2)(1 - \alpha_3 - \beta_4 - \gamma_3). \tag{37}$$

This is a fraction that must be attributed to factors not indexed in $\boldsymbol{F}$, $\mathbf{G}$.

Other fractions of the phenotype variance can be worked out from consideration of terms that appear in differences that yield the fractions. Here, we mention three that will be used in our examples. The difference between DZ twins and full siblings in correlations of residual scores is a measure of the importance of cohort effects because twins are born at the same time of the same pregnancy and full siblings are born of different pregnancies. The fraction of phenotype variance associated with this is

$$(1 - h^2)(\alpha_3 + \beta_4 + \gamma_2) - (1 - h^2)(\alpha_2 + \beta_3 + \gamma_2) = (1 - h^2)[(\alpha_3 - \alpha_2) + (\beta_4 - \beta_3)]. \tag{38}$$

The correlative effects of pre- and postnatal familial environments on twins are indicated by α_3 and β_4. If cohort effects are subtracted from these, then the remainder can be used to indicate effects of pre- and postnatal familial environments adjusted for cohort differences. These are

$$(1 - h^2)[\alpha_3 - (\alpha_3 - \alpha_2)] = (1 - h^2)\alpha_2 \text{ and } (1 - h^2)[\beta_4 - (\beta_4 - \beta_3)] = (1 - h^2)\beta_3 \tag{39}$$

respectively. Each of the fractions in (38) and (39) must be interpreted in the same manner as (36); that is, a nonzero fraction indicates nonzero effects for at least one genotype; it does not mean that the effects cannot sum out over all genotypes.

Model expressions can be used to investigate the meaning of parameters in other genetic analyses. We present four examples of these here, which will be used in our worked example. The first two are simple functions of correlations in phenotypes, which are often used to assess the importance of genetic effects. They are the doubled difference in corre-

lations of phenotype scores for MZ twins and DZ twins and the coefficient of regression of child's score on the average of parents' scores [cf 8]. In terms of our model parameters, the respective expressions for these are

$$h^2(1-\rho) - 4h(1-h^2)^{1/2}(\delta + \Delta) + 2(1-h^2)(\gamma_3 - \gamma_2) \tag{40}$$

and

$$\frac{h^2(1+\rho) + 2h(1-h^2)^{1/2}\delta + (1-h^2)(\alpha_0 + \alpha_1 + 2\beta_2 + 2\gamma_2)}{h^2(1+\rho) + (1-h^2)(1+\theta)}. \tag{41}$$

The first is obtained from lines 1 and 2 of Table 3, and the second, which is the sum of correlations between child and mother and father divided by 1 plus the correlation between husband and wife, is obtained from lines 4, 5, and 14. The expression in (40) equals h^2 or g^2 only if

$$h^2\rho + 4h(1-h^2)^{1/2}(\delta + \Delta) + 2(1-h^2)\gamma_2 = 2(1-h^2)\gamma_3 \text{ or } (1-h^2)\gamma_3. \tag{42}$$

Special cases where one or the other equality holds are easy to find; eg,

$$\rho = \delta + \Delta = 0 \text{ and } \gamma_2 = \gamma_3 \text{ or } \gamma_2 = \gamma_3/2,$$

but from data on twin correlations only, it would be impossible to tell if any of the special cases is applicable. Conditions for the second expression to equal h^2 or g^2 are more complicated, but similar conclusions hold about analyses based only on the estimated regression of a child's score on the average of parents' scores. The final two functions are the correlation for MZ twins raised apart and the deviation of it from the correlation for MZ twins raised together [cf 2]. If separation affects only the correlative effects of pre- and postnatal environments and results in the elimination of the latter, then our expressions for the correlation between twins raised apart is

$$h^2 + (1-h^2)(\alpha_3' + \gamma_3) = g^2 + (1-h^2)\alpha_3'. \tag{43}$$

In this, the prime on α_3 denotes the possibility of a modification of residual prenatal effects as a result of separation. Then the difference in correlations is

$$(1-h^2)[(\alpha_3 - \alpha_3') + \beta_3] = (1-h^2)\beta_3 + (1-h^2)(\alpha_3 - \alpha_3'). \tag{44}$$

The values of these clearly depend on α_3' and $\alpha_3 - \alpha_3'$. If $\alpha_3 = \alpha_3' = 0$, then (43) and (44) are expressions for fractions of the phenotype variance which can be attributed to variation in genetic scores and differences in postnatal familial environments. The interpretation of (43) is unchanged if only α_3' is zero, but the fraction in (44) must then be attributed to a combination of pre- and postnatal familial environments. Again, it is impossible to determine if either of these special cases applies when only correlation data on these two types of pairings are available.

STATISTICAL ANALYSIS

Several different statistical analyses are needed to make inferences about our model and parameters in it from data of the type collected in the Swedish twin-family study. The first set of analyses is used to adjust response data for coarse-grain environmental effects of sex, generation, and type of twin-family unit and to eliminate from sample correlation estimates potential biases due to the cluster-sample feature of the response data. These adjustments can be effected in two steps, the first of which involves an extension of the standard intraclass correlation analysis to account for the random-variable length of a family-unit response record, and the second is a simple procedure for combining the 72 correlation estimates into 22 summary estimates. The second set of analyses is used to obtain a test of lack of fit of the model to the 22 estimates, estimates of model parameters, and tests concerning them. In these, it is necessary to deal with problems resulting from the nonestimability of certain combinations of model parameters in general and in special cases where estimates of h^2 are near zero or one. Inequality constraints like those set out in (27) play an important part in the solution of these problems, so that explicit use of them is a unique feature of our analysis.

Inter-Unit Correlation Analysis

The 72 pair relationships can be distinguished by differences in the coarse-grain environment of Swedish-type studies and some additional differences among consanguineous and nonconsanguineous relatives. For example, a man with his DZ twin's son is characterized by two values of **F**, which denote sex of individual, generation of birth, and type of twin-family unit for each member of the pair and by the uncle-nephew relationship. Scores of a phenotype for such a subsample provide data to estimate mean values that are needed to calculate the deviations within coarse-grain environments on which all random-variable scores in (4) depend.

Pairs in a subsample are not independent. Some overlap, some are disjoint but are from the same family, and some are disjoint and from different families but are from the same twin-family unit. Therefore, ordinary sums of squares and cross-products pick up intra-family and inter-family, intra-unit variation, as well as inter-unit variation, and a special method of analyses is needed to obtain adjusted inter-unit estimates from the 72 subsamples of pairs of phenotype scores. Details of our method can be set out easily if distinctions are made between intra- and inter-family relationships and asymmetric and symmetric relationships. The first of these is obvious. For the second, symmetric will be used only for pairs in which the individuals are the same sex and are born in the same generation. Examples are the inter-family, symmetric relationships between male first cousins and the intra-family, asymmetric relationship between father and son.

Individuals and their useful identifying characteristics must be indexed in order to describe methods of calculation. We assume that this has been done and use the subscripts i, j, and k to index family units, families within a family unit, and individuals within a family. For a given pair relationship, the number of twin-family units in which at least one pair of the required type can be found is denoted by p, and values of the i index are assigned so that $1 \leqslant i \leqslant p$. The number of families in the ith unit with at least one member of a pair is denoted by q_i, and values of the j index are assigned so that $1 \leqslant j \leqslant q_i$. In symmetric relationships, the number of individuals in the jth family of the ith unit who are counted in the pairs is denoted by n_{ij}, and the values of the k index are assigned so that $1 \leqslant k \leqslant n_{ij}$. In asymmetric relationships, the numbers of individuals in the

https://doi.org/10.1017/S0001566000006590 Published online by Cambridge University Press

jth family of the ith unit who are respective first and second members of an asymmetric pair are denoted by n_{ij} and n'_{ij}. The unstandardized score of a phenotype for the kth individual in the jth family of the ith unit is denoted by x_{ijk}.

Location adjusted, unbiased inter-unit covariance estimates for all asymmetric pair relationships can be calculated from

$$1/(p-1)\left\{\left[\sum_{i=1}^{p}\sum_{j=1}^{q_i}\sum_{k=1}^{n_{ij}}\sum_{k'=1}^{n'_{ij*}}(1/m_i)x_{ijk}x_{ij*k'} - (1/p)\left(\sum_{i=1}^{p}\sum_{j=1}^{q_i}\sum_{k=1}^{n_{ij}}(m'_{ij*}/m_i)x_{ijk}\right)\times\right.\right.$$
$$\left.\left(\sum_{i=1}^{p}\sum_{j=1}^{q_i}\sum_{k'=1}^{n'_{ij*}}(m_{ij}/m_i)x_{ij*k'}\right)\right] -$$
$$(1/p)\sum_{i=1}^{p}\left[\sum_{j=1}^{q_i}\sum_{k=1}^{n_{ij}}\sum_{k'=1}^{n'_{ij*}}(1/m_i)x_{ijk}x_{ij*k'} -\right.$$
$$\left.\left.\left(\sum_{j=1}^{q_i}\sum_{k=1}^{n_{ij}}(m'_{ij*}/m_i)x_{ijk}\right)\left(\sum_{j=1}^{q_i}\sum_{k'=1}^{n'_{ij*}}(m_{ij}/m_i)x_{ij*k'}\right)\right]\right\} \qquad (45)$$

where $m_i = m_{i1}m'_{i1} + m_{i2}m'_{i2}$, $m_{ij} = n_{ij}$, $m'_{ij} = n'_{ij}$, and for respective intra-family and inter-family and inter-family pairs $j^* = j$ and $j^* = 2(1)$ if $j = 1(2)$. Two variance estimates can be obtained from the same formula by first replacing $x_{ij*k'}$ everywhere by x_{ijk} and then by reversing the order of replacement and substituting x_{ij*k} everywhere for x_{ijk}. Only slight modifications of (45) are required for symmetric pair relationships. The ratio m_{ij}/m_i must be replaced in the two positions where it appears by m'_{ij}/m_i, and for intra-family pairs, the double summation on k and k′ must be replaced by a sum over all (k, k′) pairs for which $k \neq k'$. The definitions of m_{ij}, m'_{ij}, and j^* for the respective intra-family and inter-family pairs are:

$$m_{ij} = n_{ij},\ m'_{ij} = (n_{ij} - 1),\ j^* = j, \text{ and } m_{ij} = n_{ij},\ m'_{ij} = n_{ij*},\ j^* = 2(1) \text{ if } j = 1(2). \qquad (46)$$

In Swedish-type studies, two members of a symmetric pair have the same pedigree characteristics, and therefore only one variance estimate is needed for a pairing. This can be obtained from (45), modified in the manner described, and from (46) by replacing $x_{ij*k'}$ everywhere by x_{ijk}.

Product-moment correlation ratios, $r_{h\ell}$, are formed from the covariance and variance estimates for location and scale adjusted, inter-unit estimates of the 72 intra-class correlations. For the rest of the analysis, each of these is first transformed into a Fisher z variable. For the ℓth correlation estimate in the hth summary group, this is

$$z_{h\ell} = \tfrac{1}{2}\ell n[(1 + r_{h\ell})/(1 - r_{h\ell})]\,. \qquad (47)$$

The approximate distribution of one of these is normal with expected value

$$\tfrac{1}{2}\ell_n[(1 + c_{h\ell})/(1 - c_{h\ell})] \qquad (48)$$

https://doi.org/10.1017/S0001566000006590 Published online by Cambridge University Press

where $c_{h\ell}$ is the intra-class correlation estimated by $r_{h\ell}$, and has a variance that does not exceed $1/(p_{h\ell} - 3)$. The exact variance in this approximation is always intermediate to $\frac{1}{2}(p_{h\ell} - 3)$ and $1/(p_{h\ell} - 3)$. By taking the larger limit, we can offset errors in the approximation. Average estimates for the 22 summary classes are obtained from linear combinations of the form

$$z_h = \sum_{\ell} (p_{h\ell} - 3) z_{h\ell} \Big/ \sum_{\ell} (p_{h\ell} - 3), \quad h = 1, \ldots, 22. \tag{49}$$

These are approximately normally distributed with expected values

$$\tfrac{1}{2} \ell n[(1 + c_h)/(1 - c_h)], \quad h = 1, \ldots, 22. \tag{50}$$

and variances that are proportional to

$$1 \Big/ \sum_{\ell} (p_{h\ell} - 3), \quad h = 1, \ldots, 22. \tag{51}$$

The proportionality constant is one, and z_h has minimum variance when the subsamples indexed by ℓ are independent, as in the case of husband-wife pairs. When the samples are related, as with male-male, female-female, and male-female pairs of cousins, the proportionality constant is not greater than one-half of the number of related pairs plus one, two in the example. In some of these cases in the Swedish sample, the constant is closer to one because few families have more than two children (see Table 1), and therefore only infrequently is there more than one pair of a set like male-male, female-female, and male-female in a sampled family unit. We found for the Swedish data that the proportionality constant can safely be taken as one for relationships 1, 2, 3, 14, 15, 16, 21, 22, and three-halves for the rest.

Summary correlation estimates corresponding to (49) are

$$r_h = (1 - e^{-2z_h})/(1 + e^{-2z_h}), \quad h = 1, \ldots, 22. \tag{52}$$

The units of information used in the calculation of each can be taken as the inverse of the variance of the z variable. For the Swedish study, this is either the reciprocal of (51) or two-thirds of the reciprocal.

Nonlinear Least-Squares Analysis

The sum of squared deviations,

$$Q = \sum_{h=1}^{22} I_h \left\{ z_h - \tfrac{1}{2} \ell n[(1 + c_h)/(1 - c_h)] \right\}^2 \tag{53}$$

where I_h is the number of units of information on which z_h is calculated and model expressions are substituted for c_h, can be minimized by choice of model parameters to provide estimates for the parameters and a test of fit of data to the model. This nonlinear least-squares analysis is similar in detail to one proposed and used by Rao et al [17].

https://doi.org/10.1017/S0001566000006590 Published online by Cambridge University Press

Were the z variables independently distributed, then the estimates of model parameters obtained in this manner would be efficient among estimates derived from the z_h variables, and the minimum of Q would be approximately chi square, with degrees of freedom equal to 22 minus the number of parameters estimated. The variables are undoubtedly correlated, but we have not detected any noticeable effects of such correlations in 42 applications of the analysis in our studies of the Swedish data, and therefore we postulate that they are negligible.

Only 13 functions of the 18 parameters in our model can be estimated with data on the 22 summary correlations. We use

$$\begin{aligned}
\phi_1 &= h^2 & \phi_8 &= \alpha_2 - \alpha_1 \\
\phi_2 &= \rho & \phi_9 &= (1 - h_2^2)[(\beta_1 - \beta_0) + (\gamma_1 - \gamma_0)] \\
\phi_3 &= \theta & \phi_{10} &= (1 - h_2^2)[(\beta_2 - \beta_1) + (\gamma_2 - \gamma_1)] \\
\phi_4 &= \delta & \phi_{11} &= (1 - h^2)(\beta_3 - \beta_2) \\
\phi_5 &= \Delta & \phi_{12} &= (1 - h^2)(\gamma_3 - \gamma_2) \\
\phi_6 &= (1 - h_2^2)(\alpha_0 + \beta_0 + \gamma_0) & \phi_{13} &= (1 - h^2)[(\alpha_3 - \alpha_2) + (\beta_4 - \beta_3)] \\
\phi_7 &= \alpha_1 - \alpha_0
\end{aligned} \tag{54}$$

and write the c_h in Q as expressions of these ϕ parameters, using the equations in Table 3. Values of the parameters, such as $\Delta = \alpha_0 + \beta_0 + \gamma_0 = 0$ or $\phi_4 = \phi_5 = 0$, can be specified in these expressions (such a restricted parameterization was described in Crumpacker et al [5]). The residual quadratic can then be minimized by choice of estimates for the remaining parameters subject to constraints such as (27). These are the nonlinear least-squares estimates of the model parameters, and the minimum value of Q is a lack-of-fit sum of squares. Estimated variances and covariances of the parameter estimates are elements of a matrix $\hat{\mathbf{V}} = (\hat{v}_{rs}) = (\hat{v}^{rs})^{-1}$, where $\hat{v}^{rs}$ equals

$$\sum_{h=1}^{22} I_h(\partial/\partial\phi_r)\left\{\tfrac{1}{2}\ln[(1 + c_h)/(1 - c_h)]\right\}(\partial/\partial\phi_s)\left\{\tfrac{1}{2}\ln[(1 + c_h)/(1 - c_h)]\right\} \tag{55}$$

evaluated in the nonlinear least-squares estimates of the unspecified ϕ parameters. The residual quadratic is approximately distributed as a chi-square random variable, with degrees of freedom equal to 22 minus the number of parameters estimated. In general, this is a noncentral chi square, which reduces to a central chi square when the model is adequate to describe the 22 correlations.

Linear combinations of the ϕ estimates can be used to provide estimates and estimates of bounds of meaningful functions of the model parameters. The estimated variance of one of these, such as $\hat{\psi} = \sum_r t_r\hat{\phi}_r$, is

$$\hat{v} = \sum_r \sum_s t_r t_s \hat{v}_{rs} \tag{56}$$

Significance tests can be computed from standardized ratios of the form $(\hat{\psi} - \psi_0)/\hat{v}^{1/2}$, where ψ_0 is a hypothesized value, and the approximate null distribution is the standard normal.

The parameters ρ and θ cannot be estimated with any reasonable precision in small samples if h^2 or $(1 - h^2)$ is close to zero. The reason for this is that ρ appears only in the

https://doi.org/10.1017/S0001566000006590 Published online by Cambridge University Press

model expressions in functions multiplied by h^2, and θ appears only in functions multiplied by $(1 - h^2)$. When one of the functions of h^2 is negligible, all the corresponding terms in the functions of ρ or θ in effect drop out of the model. We have not tried to interpret estimates of these parameters in the Swedish data unless $\hat{h}^2$ is at least 0.2 (for ρ) or at most 0.8 (for θ).

Interval constraints, such as $0 \leqslant h^2 \leqslant 1$ and those given in (27), and preset boundary values, such as $\delta = \Delta = 0$ and $\alpha_0 + \beta_0 + \gamma_0 = 0$, are very useful in the nonlinear least-squares analysis. First, they can be used to find bounds on certain functions of the model, which could not be estimated otherwise. This will be illustrated in the example to follow. Second, they limit the parameter space that has to be searched to find the minimum of Q. Finally, it should be noted that least-squares solutions in which some estimates are boundary values of constraint intervals, eg, $\hat{\delta} = \hat{\Delta} = 0$, do not differ from least-squares solutions in which parameters are specified to have these boundary values. This is a useful result because new analyses need not be calculated when some parameters are indicated to have such boundary values.

A set of computer programs written in FORTRAN IV is available at cost of reproduction and mailing from the authors.

ANALYSIS OF HEIGHT

Subjects in the Swedish twin-family study completed a medical questionnaire in which they were asked to state their height. All but five of 908 people provided this information. No verification of the accuracy of their reports has been made other than consistency checks at the time of coding and during computer editing of the data file. Therefore, the amount of measurement error and individual and family report bias in these data is undoubtedly greater than in data on ruled height, which might have been collected had there been sufficient interview time.

The 22 summary correlation estimates required for our analysis are presented in Table 4. These weighted averages of 72 estimates were calculated in the manner already described in order to eliminate differences due to the coarse-grain environment of the Swedish study – viz, 16 combinations of sex, generation, and sex-zygosity of family units. The entries for units of information are calculated in the manner described for our statistical analysis. Ratios in these indicate the relative amounts of information used in the calculation of correlation estimates. Notice that the relative amount of information drops as low as 9.3% (12/129) for first cousins related through DZ twin fathers (13), but that generally there is more information available to estimate the more important correlations for our analysis (1–5 and 14) than the less important ones.

Twin, full sibling, and parent-child correlation estimates from three frequently cited studies are included in Table 4 to provide a comparison for the Swedish twin-family data. These include correlations for MZ twins reared apart. Height was ruled off to the nearest one-quarter inch by volunteer student assistants for the Pearson and Lee 1903 study [15]. Allowances were made for people measured with their boots on. For the Newman, Freeman, and Holzinger 1937 study [14], ruled measurements were made by technicians in a university laboratory. It is not known exactly which of Shields' records were obtained by ruled measurement and which by self-reporting. It appears from indirect comments in his 1962 report that heights were self-reported for DZ twins and at least some of the MZ twins reared apart. Unfortunately, Shields [19] is less clear about how height data were obtained for his MZ twins reared together. Some of the response data for these pairs were taken from a booklet completed at home by the twins before office interviews. The booklet was a revision of a version used four years earlier to obtain information on the twins who were

*TABLE 4. Summary Correlation Estimates From Swedish Twin-Family, Self-Reported Height Data for the 22 Pair Relationships Shown in Table 3**

	Swedish-twin-family study (1977)		Pearson and Lee (1903)		Newman, Freeman and Holzinger (1937)		Shields (1962)			
							Females		Males	
Pair relationship	r	Units of information[a]	r	Units of information	r	Units of information	r	Units of information	r	Units of information
1.	0.78	66			0.94	47	0.94	23	0.98	15
					0.97	16	0.82	24	0.82	14
2.	0.41	56			0.65	47	0.44	15		
3.	0.36	72	0.54	1,070						
4.	0.36	127	0.50	1,213						
5.	0.47	129	0.51	1,224						
6.	0.22	37								
7.	0.40	34								
8.	0.07	27								
9.	0.30	30								
10.	0.19	29								
11.	0.19	19								
12.	0.05	17								
13.	0.45	12								
14.	0.27	126	0.28	1,076						
15.	0.26	69								
16.	0.08	56								
17.	0.15	38								
18.	0.01	34								
19.	−0.11	26								
20.	0.25	30								
21.	0.08	69								
22.	−0.07	52								

*Estimates have been adjusted for differences in sex, generation, and sex-zygosity of twin-family units and pertain to inter-unit variation in pairs. Correlation estimates from one family and two twin studies are included for comparisons. Height was ruled for Pearson and Lee (1903) and for Newman, Freeman and Holzinger (1937). At least some heights were self-reported for Shields (1962).

[a]Lower bounds on the number of units of information are given for the Swedish and Pearson and Lee studies. Exact values are given for the Newman, Freeman and Holzinger and Shields studies. The reciprocal of units of information is the approximate variance of the z-transformation of a correlation estimate, and $1 - r^2$ times a reciprocal is an estimate of the sampling variance of an r.

reared apart. The original must have contained questions about height, because data from twins reared apart who were never interviewed were used in his height analyses. If such questions were deleted in the revision and heights of the twins reared together were measured by rule during the office interview, then, as will be apparent in the following discussion, much of the discrepancy between Shields' data and the rest can be easily explained.

An immediate impression gained from a review of the estimates is that correlations for self-reported height are lower than for ruled height. The differences can easily be seen in the four studies for MZ twins reared together and for twin and nontwin full siblings. Shields' average estimate of 0.96 for MZ twins reared together is similar to the Newman et al al 0.94 estimate for ruled height. The average of these, 0.95, differs significantly from the 0.78 Swedish estimate for reported height. The 0.97 estimate from the Newman et al study for MZ twins reared apart is close to their estimate for twins reared together. Shields' 0.82 estimate of the same parameter is much closer to the 0.78 value for reported height of Swedish twins reared together. Similarly, his 0.44 estimate for reported height of DZ twins agrees well with the 0.41 value for the Swedish DZ twins and differs considerably from the Newman et al estimate of 0.65 for ruled height. The Pearson and Lee estimate of 0.54 for ruled heights of nontwin full siblings is substantially higher than the 0.36 Swedish estimate and intermediate to the combined Shields' and Swedish estimates and the Newman et al estimate for DZ twins.

For our analyses, we set $\alpha_0 + \beta_0 + \gamma_0 = 0$, principally because there is no evidence from earlier studies of appreciable correlations in nonadditive genetic and environmental scores for relatives less related than half-siblings. The levels retained for α and β index the fine-grain environment for self-reported height in the Swedish study. There are 12 combinations of these, which indicate differences in pre- and postnatal familial environments associated with comparisons between twin-family units, between families within units, and among children within families.

There are 14 parameters for the model expressions to be fit by the 22 summary correlation estimates. These are h^2, ρ, θ, δ, Δ, α_1, α_2, α_3, β_1, β_2, β_3, β_4, γ_1, γ_2 and γ_3. However, only 12 functions of these, eg, h^2, ρ, θ, δ, Δ, and $(1 - h^2)$ times α_1, α_2, $(\beta_2 + \gamma_1)$, $[(\beta_2 - \delta_1) + (\gamma_2 - \gamma_1)]$, $(\beta_3 - \beta_2)$, $(\gamma_3 - \gamma_2)$, and $[(\alpha_3 - \alpha_2) + (\beta_4 - \beta_3)]$, can be estimated. Therefore, there are $22 - 12 = 10$ degrees of freedom associated with the residual sum of squares with which lack of fit is tested. Three of the 10 degrees of freedom are associated with with model-predicted replications in the data. These are the pairs of correlations for relationships 4 and 6, 5 and 7, and 9 and 11. The remaining 7 degrees of freedom are associated with overall differences between the estimated pattern of correlations and the most closely fitting pattern predicted with the model. The value of the residual sum of squares, 8.56 (Table 5) is far from being significant ($P = 0.58$). We judge the fit of model to the data to be adequate.

Estimates of parameter functions minimize the residual sum of squares (53) subject to the constraints displayed in (27). Numerical values of the estimated functions are:

$\hat{h}^2 = 0.52$ $\quad$ $(1 - \hat{h}^2)\hat{\alpha}_2 = 0.00$

$\hat{\rho} = 0.28$ $\quad$ $(1 - \hat{h}^2)\widehat{(\beta_1 + \gamma_1)} = 0.00$

$\hat{\theta} = 0.28$ $\quad$ $(1 - \hat{h}^2)[\widehat{(\beta_2 - \beta_1)} + \widehat{(\gamma_2 - \gamma_1)}] = 0.06$

$\hat{\delta} = 0.00$ $\quad$ $(1 - \hat{h}^2)\widehat{(\beta_3 - \beta_2)} = 0.00$

$\hat{\Delta} = 0.00$ $\quad$ $(1 - \hat{h}^2)\widehat{(\gamma_3 - \gamma_2)} = 0.20$

$(1 - \hat{h}^2)\hat{\alpha}_1 = 0.00$ $\quad$ $(1 - \hat{h}^2)[\widehat{(\alpha_3 - \alpha_2)} + \widehat{(\beta_4 - \beta_3)}] = 0.01$

https://doi.org/10.1017/S0001566000006590 Published online by Cambridge University Press

*TABLE 5. Analysis of Self-Reported Height From Swedish Twin-Family Study**

Source of variance in phenotype scores	Parameter function	Estimate	Estimated standard deviation	Significance probability
Genotype score	g^2	0.72–0.78	0.19, 0.13[a]	<0.01, <0.01
Additive genetic score	h^2	0.52	0.31	0.05
Nonadditive genetic score	$(1 - h^2)\gamma_3$	0.20–0.26	0.21, 0.34	0.17, 0.22
Prenatal familial environment other than cohort differences	$(1 - h^2)\alpha_2$	0.00		
Postnatal familial environment other than cohort differences	$(1 - h^2)\beta_3$	0.00–0.06	0.16	0.35
Combined pre- and postnatal environment associated with cohort differences	$(1 - h^2)[(\alpha_3 - \alpha_2) + (\beta_4 - \beta_3)]$	0.01	0.12	0.42
Postnatal extrafamilial environment	$(1 - h^2)(1 - \alpha_3 - \beta_4 - \gamma_3)$	0.22	0.05	<0.01
Correlations due to phenotypic assortment and convergence and population stratification				
Between additive genetic scores	ρ	0.28	0.25	0.13
Between residual scores	θ	0.28	0.35	0.21
Sum of squares for lack of fit: 8.56, 10 degrees of freedom				0.58

*Data for the analysis are shown in Table 4.

[a]The unusual pattern of estimates obtained for genotype and component scores results from a substantial negative estimate of covariance between $\hat{h}^2$ and $(1 - \hat{h}^2)\hat{\gamma}_3$ estimates.

From these and use of the constraints, we obtain $0.00 \leqslant (1 - \hat{h}^2)(\widehat{\alpha_3 - \alpha_2}), (1 - \hat{h}^2) \times (\widehat{\beta_4 - \beta_3}) \leqslant 0.01$ and $0.00 \leqslant (1 - \hat{h}^2)(\widehat{\beta_2 - \beta_1}), (1 - \hat{h}^2)(\widehat{\gamma_2 - \gamma_1}) \leqslant 0.06$. These can now be combined to provide the important estimated bounds

$$0.00 \leqslant (1 - \hat{h}^2)\hat{\beta}_3 \leqslant 0.06 \text{ and } 0.20 \leqslant (1 - \hat{h}^2)\hat{\gamma}_3 \leqslant 0.26.$$

The final estimates shown in Table 5 are:

$\hat{h}^2 = 0.52$
$(1 - \hat{h}^2)\hat{\gamma}_3 = 0.20 - 0.26$
$g^2 = \hat{h}^2 + (1 - \hat{h}^2)\hat{\gamma}_3 = 0.72 - 0.78$
$(1 - \hat{h}^2)\hat{\alpha}_2 = 0.00$

$(1 - \hat{h}^2)\hat{\beta}_3 = 0.00 - 0.06$
$(1 - \hat{h}^2)[\widehat{(\alpha_3 - \alpha_2) + (\beta_4 - \beta_3)}] = 0.01$
$(1 - \hat{h}^2)[1 - \widehat{\alpha_3 - \beta_4 - \gamma_3}] = 0.22$
$\hat{\rho} = 0.28$
$\hat{\theta} = 0.28$

Estimated sampling variances of these are obtained in the manner described from the matrix $\hat{\mathbf{V}}$ [equation (55)] and variance formula (56). Only one of these is spelled out in detail here for the purpose of illustrating the technique. The estimated bounds for $(1 - \hat{h}^2)\hat{\gamma}_3$ are

$$(1 - \hat{h}^2)(\widehat{\gamma_3 - \gamma_2}) \text{ and } (1 - \hat{h}^2)\left\{(\widehat{\beta_1 + \gamma_1}) + [\widehat{(\beta_2 - \beta_1) + (\gamma_2 - \gamma_1)}] + (\widehat{\gamma_3 - \gamma_2})\right\}.$$

Therefore, the sampling variances of the bound estimates are

$$\text{var}[(1 - \hat{h}^2)(\widehat{\gamma_3 - \gamma_2})]$$

and

$$\text{var}[(1 - \hat{h}^2)(\widehat{\beta_1 + \gamma_1})] + \text{var}\left\{(1 - \hat{h}^2)[\widehat{(\beta_2 - \beta_1) + (\gamma_2 - \gamma_1)}]\right\} + \text{var}[(1 - \hat{h}^2)(\widehat{\gamma_3 - \gamma_2})] + 2\,\text{cov}\left\{(1 - \hat{h}^2)(\widehat{\beta_1 + \gamma_1}), (1 - \hat{h}^2)[\widehat{(\beta_2 - \beta_1) + (\gamma_2 - \gamma_1)}]\right\} + 2\,\text{cov}[(1 - \hat{h}^2)(\widehat{\beta_1 + \gamma_1}), (1 - \hat{h}^2)(\widehat{\gamma_3 - \gamma_2})] + 2\,\text{cov}\left\{(1 - \hat{h}^2)[\widehat{(\beta_2 - \beta_1) + (\gamma_2 - \gamma_1)}], (1 - \hat{h}^2)(\widehat{\gamma_3 - \gamma_2})\right\}$$

respectively, and the individual variance and covariance estimates to use in these are obtained from their corresponding positions in $\hat{\mathbf{V}}$.

A second analysis of the data in which $\alpha_1, \alpha_2, \beta_1, \gamma_1$, and $(\beta_3 - \beta_2)$ are set equal to zero will not result in a change in the nonzero estimates. This, as indicated, is a characteristic of the constrained nonlinear least-squares analysis. Were any parameter set equal to a value other than its numerical estimate and a new analysis calculated, then all of the remaining estimates could change in value.

One of the most instructive observations to be made about the results in Table 5 is that it is difficult to demonstrate a significant deviation from the hypothetical values when data are fit to a many-parameter model. Normalized ratios of the form $(\hat{\psi} - \psi_0)/\hat{v}^{1/2}$ can be used to calculate approximate significance probability for deviations of estimates, $\hat{\psi}$, from hypothetical values, ψ_0. These values are shown in Table 5 for postulated zero values of the parameters. Only four of the probabilites are sufficiently small to be regarded as significant, even though one might consider all eight estimates of bounds of h^2, $(1 - h^2)\gamma_3$, g^2, $(1 - h^2)(1 - \alpha_3 - \beta_4 - \gamma_3)$, ρ, and θ to be substantially greater than zero hypothetical values.

https://doi.org/10.1017/S0001566000006590 Published online by Cambridge University Press

There are three interesting results in our analysis of the Swedish data that can be used to explain self-reported height and to compare techniques for estimating h^2, g^2, and $(1 - h^2)\beta_3$. The first is that familial environment seems to be an unimportant determinant of the phenotype because all of the variance not assigned to s(**G**) (ie, 22%) can be attributed to $\hat{v}(\boldsymbol{F}, \mathbf{G})$. At least part of this can reasonably be ascribed to reporting bias if the substantial estimate of θ is accepted as an indication of nonzero correlation in residual scores. A plausible hypothesis is that people tend to round estimates of their height up or down to agree more closely to preconceived notions of what it should be, for example similar to a popular norm or to a spouse's stature. Most of the remainder of the 22% is probably due to reporting errors, because the correlations in ruled height for MZ twins is so high ($\bar{r} = 0.95$ in Table 4), and therefore there is little variance left over (5%) to attribute to environmental differences not captured in our $\boldsymbol{F}$ scheme. The second result is the fractional division of $\hat{g}^2$ into $\hat{h}^2$ and $(1 - \hat{h}^2)\hat{\gamma}_3$. We estimate that from 67% (0.52/0.78) to 72% (0.52/0.72) of the variance in genotype scores is due to the additive genetic component. Fisher [9], in an analysis of the Pearson and Lee data, estimated that the variance of additive genetic scores for measured height in England makes up 79% of the variance in genotype scores. The third result pertains to the portion of the additive genetic fraction that can be attributed to phenotypic assortative mating and population stratification. This is $\{\text{var}[\text{t}(\mathbf{G})] - 2\,\text{var}[\text{t}(\mathbf{g})]\}/\text{var}[\text{t}(\mathbf{G})] = \text{m}/(1 + \text{m}) = \rho/2$ and is estimated to be 0.14, or 14%, in our analysis. Fisher obtained a substantially larger estimate of 27% in his analysis of the Pearson and Lee data. However, Fisher's method must be in error because, based on the rest of his analysis, the correlation between husband and wife (0.28 in Table 4) is an estimate of $h^2\rho + (1 - h^2)\theta$ (line 14, Table 3), which he estimates to be $0.79\rho + 0.21\theta$, and therefore the only admissible estimates for ρ are between (0.28–0.21)/0.79 = 0.09 and 0.28/0.79 = 0.35. The estimated percent of the additive genetic fraction due to assortative mating or population stratification should therefore be between 4% and 18%, which is in close agreement with our 14% result.

The heritability estimates based on the difference between correlations for MZ and DZ twins that can be calculated from the Newman et al data and the Shields' data are very different. The first, 0.58, is close to our h^2 estimate, while the Shields' value, 1.00, reflects the suspected difference in collection techniques in his data. The twin-difference estimate calculated from the Swedish data is 0.74, and predicted by our equation (40) evaluated in our $\hat{h}^2$, $\hat{\rho}$, $\hat{\delta}$, Δ, and $(\gamma_3 - \gamma_2)$, is 0.77. These are considerably higher than our 0.52 model estimate for h^2 and the Newman et al estimate. The first of these comparisons points out that the twin difference estimate can be considerably biased for h^2, and the second suggests that the Newman et al correlation for DZ twins is an overestimate.

Fisher used the regression coefficient of child's height on the average of parents' height to obtain his estimate of 0.79 for h^2. The same estimate from the Swedish data is (0.36 + 0.47)/(1.00 + 0.20) = 0.69 and from our modelled expression for the regression ratio [equation (41)], and the result of our analysis of the Swedish data is 0.62. The difference between 0.69 or 0.62 and our 0.52 model estimate is fairly large. If these three estimates are multiplied by a factor 1.21, the ratio of correlations for MZ twins reared apart for the Newman et al and Swedish studies, to adjust for reporting errors, which appear to account for much of the environmental difference between reported and ruled height, then the first two of the resulting values 0.83, 0.75, and 0.63 are in good agreement with Fisher's estimate. However, they and Fisher's values appear to be overestimates of h^2, which is better approximated by our adjusted model estimate, 0.63.

If the conclusion from our analysis, that $\alpha_3 \sim 0$, can be applied to the twin data collected by Newman et al and by Shields, then the correlations for MZ twins raised apart are consistent estimates for g^2 [equation (43)]. This results in an estimate of 0.97 for ruled height and, assuming that Shields used only booklet data, 0.82 for reported height. The latter is consistent with our upper bound estimate of 0.78, and the former agrees with Shields' and our estimates after they have been adjusted upwards by 1.21 to 0.97 and 0.94 to eliminate reporting errors.

REFERENCES

1. Boyle CR, Elston RC (1979): Multifactorial genetic models for quantitative traits in humans. Biometrics 35:55–68.
2. Cavalli-Sforza LL, Bodmer WF (1971): "The Genetics of Human Populations." San Francisco: W. H. Freeman and Company.
3. Cederlöf R (1966): The twin method in epidemiological studies on chronic diseases. (Doctoral Dissertation, Academick avhandling, Stockholm universitet, Stockholm).
4. Comrey AL (1970): "Comrey Personality Scales Manual." San Diego: Educational and Industrial Testing Service.
5. Crumpacker DW, Cederlöf F, Friberg L, Kimberling WJ, Sörenson S, Vandenberg SG, Williams JS, McClearn GE, Grever B, Iyer H, Krier MJ, Pedersen NL, Price RA, Roulette I (1980): A twin methodology for the study of genetic and environmental control of variation in human smoking behavior. Acta Genet Med Gemellol 28:173–195.
6. Elston RC, Rao DC (1978): Statistical modeling and analysis in human genetics. Annu Rev Biophys Eng 7:253–286.
7. Eysenck HJ, Eysenck SGB (1975): "Manual of the Eysenck Personality Questionnaire." London: Hodder and Stoughton.
8. Falconer DS (1964): "Introduction to Quantitative Genetics," Ed 2. New York: The Ronald Press Company.
9. Fisher RA (1918): The correlation between relatives on the supposition of mendelian inheritance. Trans R Soc (Edinb) 42:321–341.
10. Floderus B (1974): Psychosocial factors in relation to coronary heart disease and associated risk factors. Nord Hyg Tidskr Suppl 6.
11. Kempthorne O (1957): "An Introduction to Genetic Statistics." Ames, Iowa: Iowa State University Press.
12. Medlund P, Cederlöf R, Floderus B, Friberg L, Sörensen S (1976): A new Swedish twin registry. Acta Medica Scandinavia. Supplement 660.
13. Morton NE (1974): Analysis of family resemblance. I. Introduction. Am J Hum Genet 26:318–330.
14. Newman HH, Freeman FN, Holzinger KJ (1937): "Twins: A Study of Heredity and Environment." Chicago: University of Chicago Press.
15. Pearson K, Lee A (1903): On the laws of inheritance in man. I. Inheritance of physical characters. Biometrika 2:357–462.
16. Pedersen NL, McClearn GE (1980): "Factor Structure of Common Drug Usage." Progress Report. Institute for Behavioral Genetics, Boulder.
17. Rao DC, Morton NE, Yee S (1974): Analysis of family resemblance. II. A linear model for family correlations. Am J Hum Genet 26:331–359.
18. Russell MAH, Peto J, Patel US (1974): The classification of smoking by factorial structure of motives. J R Statis Soc A 137:313–333.
19. Shields J (1962): "Monozygotic Twins Brought Up Apart and Brought Up Together." London: Oxford University Press.

20. Vandenberg SG, Price RA (1978): Replication of the factor structure of the Comrey Personality Scales. Psychol Rep 42:343–352.
21. Williams JS, Crumpacker DW, Krier MJ (1980): Stability for a factor-analytic description of smoking behavior. Drug Alcohol Dependence (in press).

Correspondence: Professor James S. Williams, Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA.