Behavior Genetics, Vol. 32, No. 1, January 2002 (© 2002)

The Mixed or Multilevel Model for Behavior Genetic Analysis

Guang Guo[1,3] and Jianmin Wang[2]

Received 26 June 1999—Final 31 Aug. 2001

We propose the mixed model or multilevel model as a general alternative approach to existing behavior genetic analysis—an alternative to correlation analysis, the DeFries-Fulker analysis, and structural equation modeling. The mixed or multilevel model handles readily families of behavioral genetic data, which include paired sibling data (e.g., pairs of MZ and DZ twins) and clustered sibling data (e.g., a family of more than two biological siblings) as special cases. Not only can a family of behavioral genetic data have more than two siblings, it can also contain multiple types of siblings (e.g., a pair of MZ twins, a pair of DZ twins, a full sibling, and a half sibling). In contrast to the traditional approaches, the mixed or multilevel model is insensitive to the order of the siblings in a sibling cluster. We apply our approach to a large, nationally representative behavior genetic sample collected recently by the Add Health Study. We demonstrate the approach through several applications using both clustered and family complex behavioral genetic data: conventional variance decomposition analysis, analysis of interactions between genetic and environmental influences, and analysis of the possible genetic basis for friendship selection. We compare results from the mixed or multilevel model, Pearson's correlation analysis, and the structural equation model.

KEY WORDS: Multilevel model; hierarchical linear model; the mixed model; DF analysis; and structural equation models.

INTRODUCTION

There is a rapidly growing body of evidence pointing to an important part of genetics in the determination of human pathology, psychopathology, and physical traits. For example, many conditions of serious psychopathology have been found to run in families. Numerous studies show that the risk of schizophrenia for an offspring of a schizophrenic parent is about 13 times as high as that in the general population (Gottesman, 1991). The risk of manic-depressive psychosis for those with a manic-depressive parent is about 10 times as high as that for those without such a parent (Tsuang and Faraone, 1990). Consistent with our direct observation of the resemblance in physical traits among biologically related relatives, genetic studies have found that about 70–90% of variance in height and weight is attributable to genetic influences (Grilo and Pogue-Geile, 1991).

Compared with pathology and physical traits, complex human behavior appears much more environmentally determined; nevertheless, behavior geneticists have reported genetic effects on such seemingly environmentally determined behavior as parenting style, rate of accident occurrence in childhood, television viewing habits, peer groups selection, social support, marital disruption, education attainment, and socioeconomic status (Plomin *et al.*, 1994; Rowe, 1994).

In this paper, we propose the mixed (Searle *et al.*, 1992) or multilevel model (Mason *et al.*, 1983; Goldstein, 1987, 1995; Bryk and Raudenbush, 1992) as an

[1] Department of Sociology, University of North Carolina at Chapel Hill, NC.
[2] Department of Biostatistics, University of North Carolina at Chapel Hill, NC.
[3] To whom correspondence should be addressed at Department of Sociology, University of North Carolina at Chapel Hill, 27599.

0001-8244/02/0100-0037/0

alternative statistical tool for behavior genetic analysis—an alternative to the traditional correlation analysis, the DeFries-Fulker analysis (DeFries and Fulker, 1985; Rodgers and McGue, 1994; Cherny *et al.*, 1992), and structural equation models (Neale and Cardon, 1992). When explicit measures of genes are unavailable, behavior genetic analysis resorts to genetically related individuals clustered into families. In this paper, we show that the mixed model readily handles complex sibling structure. Treating individuals as level-one units and families as level-two units, the mixed model can be conceptualized as a two-level multilevel linear model. While most of our presentation is in terms of the mixed model, we will make a point of relating it to the multilevel model because of the popularity of the latter among social scientists.

We classify behavior genetic data into three types. The first consists of paired sibling data such as MZ or DZ twins. In this type of data, while the individuals within a pair are correlated due to genetic relatedness, the individuals across pairs are considered independent. We refer to the second type of data as clustered behavior genetic data, with each cluster containing two or more siblings. The number of siblings in a cluster may or may not be the same across clusters. All pairs of siblings within a cluster are genetically related in the same way. A cluster of four full siblings is an example of clustered behavior genetic data. Paired sibling data can be viewed as clustered data in which every cluster consists of two genetically related individuals. Families of behavior genetic data are the third and the most general type of behavior genetic data. Some families are not only larger, containing three or more siblings, but also more complicated genetically, containing more than one type of sibling. Conceivably, a family may consist of a pair of MZ twins, a pair of DZ twins, two full siblings, a half sibling, and two step siblings. Correlation analysis and the DF analysis appear unable to handle clustered or families of behavior genetic data. Structural equation models are a much more general and flexible approach, though the programming required for analyzing family data is likely to be complex. In the next section, we describe the mixed model. Then we show how the mixed or multilevel linear model can be adapted to analyze clustered behavior genetic data, and we describe the mixed or multilevel model for the more complex families of behavior genetic data. In both of these sections, we demonstrate the methodology for behavior genetic analysis using a nationally representative genetically informative sample collected by the Add Health Study (Bearman *et al.*, 1997). In the final section, we offer concluding remarks.

THE MIXED MODEL

The general form of the mixed model is typically described as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{Y}$ is the vector of observed Y_is, $\boldsymbol{X}$ is the matrix of observed predictors, $\boldsymbol{\beta}$ is the vector of parameter for the observed predictors, $\boldsymbol{Z}$ is the known design matrix for the vector of unknown random effects $\boldsymbol{u}$, and $\boldsymbol{e}$ is the vector of random errors, e_i's (Searle, 1971; Searle *et al.*, 1992). The mixed model assumes that $\boldsymbol{u}$ and $\boldsymbol{e}$ are mutually independent and each normally distributed with $E[\boldsymbol{u}] = \boldsymbol{0}$, $E[\boldsymbol{e}] = \boldsymbol{0}$, $\mathrm{Var}[\boldsymbol{u}] = \boldsymbol{G}$, and $\mathrm{Var}[\boldsymbol{e}] = \boldsymbol{R}$. Then the covariance matrix of $\boldsymbol{Y}$ is $\mathrm{Var}[\boldsymbol{Y}] = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z} + \boldsymbol{R}$. The parameters in $\boldsymbol{G}$ and $\boldsymbol{R}$ can be estimated by the method of maximum likelihood (ML) or restricted maximum likelihood (REML). With $\boldsymbol{G}$ and $\boldsymbol{R}$, $\boldsymbol{\beta}$ can be estimated via the generalized least squares. In the following sections, we will show how the general form of the mixed model (1) can be applied to specific problems in behavior genetic studies. Earlier work by Eaves and Gale (1974) and Eaves *et al.* (1978) for behavior genetic analysis is closely related to, and may be considered a special case of, the mixed model.

MODELS FOR CLUSTERED BEHAVIORAL GENETIC DATA

The Model

Clustered genetically informative data contain two or more of the following types of genetically related individuals: monozygotic twins, dizygotic twins, full siblings, half siblings, and cousins. Each cluster contains only one type of genetically related individual and the individuals across clusters are independent. Treating individuals as level-one units and clusters as level-two units, the following two-level model without covariates can be used to calculate the correlation for a sample of genetically related individuals of a single type:

$$Y_{ij} = \beta_0 + u_j + e_{ij} \tag{2}$$

where Y_{ij} is the observed linear outcome for individual i in cluster j, β_0 is the intercept, u_j is the cluster-specific random effect, and e_{ij} is the individual-specific random effect or the OLS-like error term. The standard assumptions are that u_j and e_{ij} are mutually independent $N(0, s_u^2)$ and $N(0, s_e^2)$ random variables. The within-

cluster or intraclass correlation can be obtained from $\rho = \sigma_u^2/(\sigma_u^2 + \sigma_e^2)$. When a sample contains pairs of MZ twins, ρ would be the correlation between the twins.

Model (2) is a special case of the general mixed model (1) and the matrices for (2) can be written as $\boldsymbol{Y} = (Y_{11}, Y_{21}, \ldots, Y_{n_1 1}, Y_{12}, \ldots, Y_{n_N N})'$, where N is the total number of clusters in the sample, and the total number of elements in this vector is $M = \Sigma_{j=1}^{N} n_j$:

$$\boldsymbol{X} = 1_M;$$

$$\beta = \beta_0;$$

$$\boldsymbol{Z}_j = \mathbf{1}_{n_j \times 1}, j = 1, \ldots, N$$

$$\boldsymbol{Z} = \begin{pmatrix} Z_1 & & & \\ & Z_2 & & \\ & & 0 & \\ & & & Z_N \end{pmatrix};$$

$$\boldsymbol{u} = (u_1, u_2, \ldots, u_N);$$

$$\boldsymbol{e} = (e_{11}, e_{21}, \ldots, e_{n_1 1}, e_{12}, \ldots, e_{n_N N})'$$

$$\boldsymbol{G} = s_u^2 I_N \quad \text{and} \quad \boldsymbol{R} = s_e^2 I_M$$

In the tradition of multilevel models, model (2) is referred to as the combined model and the same model is frequently represented by an equation system:

$$\begin{aligned} Y_{ij} &= \beta_{0j} + e_{ij} \quad \text{(level 1 model)} \\ \beta_{0j} &= \beta_0 + u_j \quad \text{(level 2 model)} \end{aligned} \tag{3}$$

To cope with multiple types of clusters (MZ twins, DZ twins, full siblings, half siblings, and cousins) and to incorporate environmental influences, we expand (2) to

$$Y_{ij(t)} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \ldots + \beta_P x_{Pij} + u_{j(t)} + e_{ij(t)} \tag{4}$$

where $t = m, d, f, h,$ or c, indicating *type* of genetic relatedness within the clusters of individuals, and p indexes P number of environmental variables, x_{ij}s. The SAS codes for model (4) are given in Appendix 1. Allowing the variance of e_{ij} to vary by genetic relatedness (t) is crucial because the genetic theory expects the within-cluster variance for more genetically related clusters to be smaller. Following the standard practice in behavior genetic analysis, we treat full siblings and DZ twins separately even though they have the same genetic relatedness. For a sample comprising multiple types of clusters, the within-cluster correlation is $\rho_{(t)} = \sigma_{u(t)}^2/(\sigma_{u(t)}^2 + \sigma_{e(t)}^2)$ and the within-cluster correlation for MZ twins is $\rho_{(m)} = \sigma_{u(m)}^2/(\sigma_{u(m)}^2 + \sigma_{e(m)}^2)$.

Model (4) is also a special case of the general mixed model (1) and we give the matrices for (4) that differ from the matrices for (2) as follows:

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{111} & \ldots & x_{p11} \\ \mathsf{M} & \mathsf{M} & \ldots & \mathsf{M} \\ 1 & x_{1n_1 1} & \ldots & x_{pn_1 1} \\ 1 & x_{112} & \ldots & x_{p12} \\ \mathsf{M} & \mathsf{M} & \ldots & \mathsf{M} \\ 1 & x_{1n_N N} & \ldots & x_{pn_N N} \end{pmatrix};$$

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p);$$

$\boldsymbol{Z}_j = \mathbf{1}_{nj \times 1}(z_{j(m)}, z_{j(d)}, z_{j(f)}, z_{j(h)}, z_{j(c)})$, where $z_{j(i)}$ is an indicator variable taking the value of 1 if the cluster type is t, and 0 otherwise:

$$\boldsymbol{u} = (u_{1(m)}, u_{1(d)}, u_{1(f)}, u_{1(h)}, u_{1(c)}, \ldots, u_{N(m)}, \ldots, \boldsymbol{u}_{N(c)})';$$

$$\boldsymbol{G} = \boldsymbol{I}_N \otimes \begin{pmatrix} \sigma_{u(m)}^2 & & & \\ & \sigma_{u(d)}^2 & & \\ & & 0 & \\ & & & \sigma_{u(c)}^2 \end{pmatrix}; \quad \text{and}$$

$$\boldsymbol{R} = \begin{pmatrix} r_1 & & & \\ & r_2 & & \\ & & 0 & \\ & & & r_M \end{pmatrix}$$

where $r_k = \Sigma_t z_{k(t)} \sigma_{e(t)}^2$.

Again, level-1 and level-2 models in (4) can be written separately as in the multilevel model literature:

$$\begin{aligned} Y_{ij(t)} &= \beta_{0j(t)} + \beta_1 x_{1ij} + \quad \text{(level 1 model)} \\ &\quad \beta_2 x_{2ij} + \ldots + \beta_p x_{pij} + e_{ij(t)} \\ \beta_{0j(t)} &= \beta_0 + u_{j(t)} \quad \text{(level 2 model)} \end{aligned} \tag{5}$$

The results from model (4) or (5) can be used to compute the within-cluster correlation by type of cluster with or without adjusting for environmental influences. To compute the correlations, we construct the following system equations under the usual assumptions in behavior genetic studies, including additive genetic variance, little or no assortative mating, and equal shared environmental influences across different types of clusters

$$\begin{aligned} h_x^2 + c_{(md),x}^2 &= \rho_{(m),x} \\ (1/2)\, h_x^2 + c_{(md),x}^2 &= \rho_{(d),x} \\ (1/2)\, h_x^2 + c_{(f),x}^2 &= \rho_{(f),x} \\ (1/4)\, h_x^2 + c_{(h),x}^2 &= \rho_{(h),x} \\ (1/8)\, h_x^2 + c_{(c),x}^2 &= \rho_{(c),x} \end{aligned} \tag{6}$$

where h_x^2 is the heritability in the environment described by x and $c_{(md),x}^2, c_{(f),x}^2, c_{(h),x}^2$, and $c_{(c),x}^2$ are the proportions of the variance owing to shared environmental influences among MZ and DZ twins, full siblings, half siblings, and cousins, respectively, also in the environment described by x. The first equation in (6) holds because the total correlation within MZ twin pairs must be equal to the sum of the correlation due to heritability and the correlation due to the shared environmental influences. Given the results from (4) or (5), (6) has five equations and five unknowns and can be solved for $h_x^2, c_{(md),x}^2, c_{(f),x}^2, c_{(h),x}^2$, and $c_{(c),x}^2$. Note that the proportion of the variance due to shared environmental influences is allowed to vary across twins, full siblings, half siblings, and cousins. When a single shared environmental parameter, c_x^2, is preferred, we have an overidentified model. In such a case, the single shared environmental parameter c_x^2 may be obtained by the least-squares method, in which the estimated correlations are regressed on the coefficients (1, 1/2, 1/2, 1/4, and 1/8) of h_x^2. The intercept in such a regression would be c_x^2.

Model (4) can be expanded to incorporate the random effects of environmental influences

$$Y_{ij(t)} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \ldots + \beta_p x_{pij} + u_{1j} x_{1ij} + u_{2j} x_{2ij} + \ldots + u_{pj} x_{pij} + u_{0j(t)} + e_{0ij(t)} \quad (7)$$

where x_{pij} is the pth environmental influence for individual i in cluster j; β_0 is the intercept; $\beta_1, \beta_2, \ldots, \beta_p$ and $u_{1j}, u_{2j}, \ldots, u_{pj}$ are the fixed and random coefficients of the environmental influences, respectively; $u_{0j(t)}$ and $e_{0ij(t)}$ are the random effects at the cluster and individual levels, respectively; and $\mathrm{Var}(u_{0j(t)}) = \sigma_{u0(t)}^2$, $\mathrm{Var}(u_{1j}) = \sigma_{u1}^2$, , . . . , $\mathrm{Var}(u_{pj}) = \sigma_{uP}^2$, $\mathrm{Var}(e_{0ij(t)}) = \sigma_{e0(t)}^2$. We provide SAS codes for model (7) in Appendix 2.

Model (7) is again an application of the general mixed model (1). In the following, we describe the matrices for (7) that have not been described before:

$$\boldsymbol{Z}_j = \begin{pmatrix} x_{11j} & \ldots & x_{P1j} & z_{j(m)} & z_{j(d)} & z_{j(f)} & z_{j(h)} & z_{j(c)} \\ \mathsf{M} & \ldots & \mathsf{M} & \mathsf{M} & \mathsf{M} & \mathsf{M} & \mathsf{M} & \mathsf{M} \\ x_{1n_jj} & \ldots & x_{Pn_jj} & z_{j(m)} & z_{j(d)} & z_{j(f)} & z_{j(h)} & z_{j(c)} \end{pmatrix};$$

$\boldsymbol{u} = (u_{11}, u_{21}, \ldots, u_{P1}, u_{01(m)}, u_{01(d)}, \ldots, u_{01(c)}, \ldots, u_{1N}, u_{2N}, \ldots, u_{PN}, u_{0N(m)}, u_{0N(d)}, \ldots, u_{0N(c)})'$; and

$$\boldsymbol{G} = \boldsymbol{I}_N \otimes \begin{pmatrix} \sigma_{u1}^2 & & & & & & \\ & 0 & & & & & \\ & & \sigma_{uP}^2 & & & & \\ & & & \sigma_{u0(m)}^2 & & & \\ & & & & \sigma_{u0(d)}^2 & & \\ & & & & & 0 & \\ & & & & & & \sigma_{u0(c)}^2 \end{pmatrix}$$

Viewing (7) as a combined multilevel model, we can write level-1 and level-2 models in (7) separately as

$$\begin{aligned} Y_{ij(t)} &= \beta_{0j(t)} + \beta_{1j} x_{1ij} + \beta_{2j} x_{2ij} + \ldots + \beta_{Pj} x_{Pij} + e_{ij(t)} && \text{(level 1 model)} \\ \beta_{0j(t)} &= \beta_0 + u_{0j(t)} && \text{(level 2 model)} \\ \beta_{1j(t)} &= \beta_1 + u_{1j(t)} && \text{(level 2 model)} \\ \beta_{2j(t)} &= \beta_2 + u_{2j(t)} \mathsf{K} && \text{(level 2 model)} \\ \beta_{pj(t)} &= \beta_p + u_{pj(t)} && \text{(level 2 model)} \end{aligned} \quad (8)$$

A more complicated model may be obtained by adding one or more environmental influences in one or more of the level-2 models in (8). Model (7) can be used to study the interaction between heritability and environment or the influences of nurture on the expression of nature. When environmental influences such as family income and parental education do not vary within a cluster, x_{pij} in (7) or (8) simplifies to x_{pj} for all xs. Assuming $\mathrm{Cov}(u_{pj}, u_{p'j}) = 0$ for $p \neq p'$, we have

$$\rho_{(t)}(Y_{ij}, Y_{i'j} \mid_j, |x_j) = \frac{\sigma_{u0(t)}^2 + \sigma_{u1}^2 x_{1j}^2 + \ldots + \sigma_{uP}^2 x_{Pj}^2}{\sigma_{u0(t)}^2 + \sigma_{e0(t)}^2 + \sigma_{u1}^2 x_{1j}^2 + \ldots + \sigma_{uP}^2 x_{Pj}^2} \quad (9)$$

Now the within-cluster correlation, $\rho_{(t)}(Y_{ij}, Y_{i'j} | \boldsymbol{x}_j)$, is a function of both the type of genetic relatedness and environmental influences. Given $\rho_{(t)}(Y_{ij}, Y_{i'j} | \boldsymbol{x}_j)$ for each t and using (6), we can calculate the level of genetic or shared environmental influences for each measured environment.

When measures of environmental influences are available at the contextual level (neighborhood or school), model (7) or (8) can be expanded further in the framework of multilevel models:

$$Y_{ijk(t)} = \beta_0 + \beta_1 x_{ijk} + \gamma_1 z_k = u_{1jk} x_{ijk} + v_{1k} z_k + v_{0k} + u_{0jk(t)} +, e_{0ijk(t)} \quad (10)$$

where x_{ijk} and z_k are measures of environmental influences at the individual and contextual levels, respectively; β_1 and γ_1 are their fixed effects; v_{0k}, $u_{0jk(g)}$, and $e_{0ijk(g)}$ are random intercepts at the individual, cluster, and contextual levels, respectively; u_{1jk} and v_{1k} are the random coefficients for x_{ijk} and z_k, respectively. The covariances across individual, cluster, and contextual levels are assumed to be zero. Model (10) is a three-level multilevel model with individuals as level-1 units, clusters as level-2 units, and contexts as level-3 units.

Alternatively, model (10) can be described as a multiple equation system:

$$\begin{aligned} Y_{ijk(t)} &= \beta_{0jk(t)} + \beta_{1jk} x_{ijk} && \text{(level 1 model)} \\ &\quad + \gamma_{1k} z_k + e_{0ijk(t)} \\ \beta_{0jk(t)} &= \beta_{0k} + u_{0jk(t)} && \text{(level 2 model)} \\ \beta_{1jk} &= \beta_1 + u_{1jk} && \text{(level 2 model)} \\ \beta_{0k} &= \beta_0 + v_{0k} && \text{(level 3 model)} \\ \gamma_{1k} &= \gamma_1 + v_{1k} && \text{(level 3 model)} \end{aligned} \tag{11}$$

For simplicity, the model has only two environmental variables, one at the individual level and the other at the contextual level. Adding more environmental variables at either or both levels is straightforward.

Applications of the Models

Application Data. In this section, we present results from applications of the mixed or multilevel model. The data sources for these applications are the National Longitudinal Study of Adolescent Health (Add Health). Add Health is a school-based study of the health-related behaviors of adolescents in grades 7–12 in the United States. In 1994, the in-school questionnaire was completed by more than 90,000 adolescents from 134 schools. All students who completed an in-school questionnaire as well as those who did not complete a questionnaire but who were listed on a school roster were eligible for selection into the in-home sample. A total of 12,105 adolescents were actually interviewed in the first wave of the in-home survey from April through December of 1995.

The data for our analysis come from the kinship sample within the Add Health Study, which has deliberately incorporated the behavior-genetic designs as components in an otherwise traditional survey. The genetically informative sample is composed of six groups: MZ twins, DZ twins, full biological siblings, half biological siblings, cousins, and biologically unrelated adolescents living in the same household. Having screened a population of more than 90,000 adolescents, Add Health has identified a large number of genetically related individuals. With six degrees of genetic relatedness embedded in a nationally representative survey, this Add Health genetic sample provides a unique opportunity to examine the contribution of genetic and environmental influences on children's well-being. We use two measure for children's well-being: intellectual development measured by an abridged version of the Peabody Picture Vocabulary Test-Revised (PPVT) and grade point average (GPA).

Application I: Variance Decomposition Analysis. Table I presents parameter estimates and their standard errors of three mixed or multilevel models of intellectual development as examples of correlation analysis. All three models can be viewed as special cases of

Table I. Parameter Estimates and Their Standard Errors of the Mixed Models of Intellectual Development as Examples of Correlation Analysis: Model A Allowing a Single Across-Cluster Variance, Model B Allowing a Single Across-Cluster Variance and Controlling for Race, and Model C Allowing Across-Cluster Variance to Vary by Type of Genetic Relatedness and Controlling for Race

	Model A Parameter (se)	Model B Parameter (se)	Model C Parameter (se)
Intercept	100.6 (0.30)	104.1 (0.317)	104.0 (0.314)
Black		−12.0 (0.584)	−12.2 (0.565)
σ^2_u, (for all types)	95.60 (5.13)	66.08 (4.12)	
$\sigma^2_{u(m)}$ (MZ twins)			117.05 (14.9)
$\sigma^2_{u(d)}$ (DZ twins)			57.97 (8.33)
$\sigma^2_{u(f)}$ (full sibs)			60.22 (6.08)
$\sigma^2_{u(h)}$ (half sibs)			54.90 (10.3)
$\sigma^2_{u(c)}$ (cousins)			39.14 (16.7)
$\sigma^2_{e(m)}$ (MZ twins)	48.41 (5.45)	52.27 (6.14)	43.84 (4.63)
$\sigma^2_{e(d)}$ (DZ twins)	87.01 (6.22)	86.27 (6.02)	87.53 (6.58)
$\sigma^2_{e(f)}$ (full sibs)	87.66 (4.42)	87.88 (4.38)	87.91 (4.67)
$\sigma^2_{e(h)}$ (half sibs)	94.19 (8.07)	92.93 (7.73)	95.56 (8.74)
$\sigma^2_{e(c)}$ (cousins)	132.3 (18.5)	115.4 (15.0)	121.6 (17.9)
−2 logLikelihood	24591.7	24217.1	24189.5
Sample size	3129	3129	3129

model (4). Model A allows a single across-cluster variance ($\sigma^2_{u(t)} = \sigma^2_u$); model B is the same as the first model except that it controls for race; and model C is the same as model B except that it allows the across-cluster variance to vary by genetic relatedness ($\sigma^2_{u(t)} = \sigma^2_{u(t)}$). By the likelihood ratio test, model (2) is highly significant against model A, and model C is highly significant against model B.

Table II presents correlation coefficients for intellectual development by genetic relatedness. Column (a) presents the correlations estimated by Pearson's correlation analysis using five race-adjusted samples each for one type of genetic relatedness. The five race-adjusted samples are residuals from a simple linear regression with race as a single predictor (a black/white dummy variable). The estimates indicate that the strength of correlation varies systematically by genetic relatedness, and these results are consistent with the well-known findings reported previously (Plomin *et al.*, 1994). Columns b, c, and d are calculated from models A, B, and C of Table I, respectively. These results indicate that varying between-cluster variance by genetic relatedness is necessary for replicating the correlation estimates traditionally obtained through correlation analysis (column a vs. column d). In this particular analysis, varying between-cluster variance can be justified statistically by a likelihood ratio test. These results also suggest that traditional correlation analysis always allows the within-as well as between-cluster variance to vary by genetic relatedness.

Application II: Genetic Basis for Friendship Selection Among Adolescents. Friendship is built more on similarity than differences. Friends are found to be similar in height (Berkowitz, 1969), activities, needs, attitudes, and personality (Berscheid and Walster, 1978). It is hypothesized that part of the similarity is genetic.

Table II. Correlation Coefficients Estimated by (a) Conventional Correlation Analysis Using Separate Samples by Type of Genetic Relatedness, (b) Based on Model A in Table I, (c) Based on Model B in Table I, and (d) Based on Model C in Table I

Type of genetic relatedness	Correlation coefficient (number of pairs)			
	a	b	c	d
MZ twins	.727 (179)	.664	.558	.727
DZ twins	.399 (354)	.524	.433	.398
Full siblings	.399 (726)	.522	.429	.406
Half siblings	.350 (243)	.504	.415	.364
Cousins	.249 (89)	.419	.364	.243

Friends are expected to be more genetically similar to each other than randomly paired individuals.

Figure 1 illustrates a test of the hypothesis. Figure 1 shows the relationships among four persons: person A, A's friend, person B, and B's friend. Person A and A's friend are likely to be similar with respect to such personal characteristics as intellectual development and certain attitudes; so are person B and B's friend. The friendship such as that between A and A's friend or B and B's friend, however, does not clue us in on whether any of the similarity between the two friends is genetic.

To test the hypothesis of whether friendship has a genetic basis, we let A and B be pairs of MZ twins, DZ twins, full siblings, half siblings, and cousins with different and known genetic relatedness. If individuals do select friends partially according to their genetic propensity, we should see a positive correspondence between the extent to which A and B are genetically related and the strength of similarity between A's friend and B's friend. For instance, we should find friends of MZ twins significantly more similar than friends of DZ twins, and friends of DZ twins about equally similar to friends of full siblings. The inclusion of MZ twins and DZ twins in the same analysis is necessary for obtaining a genetic interpretation for friendship selection because the sim-

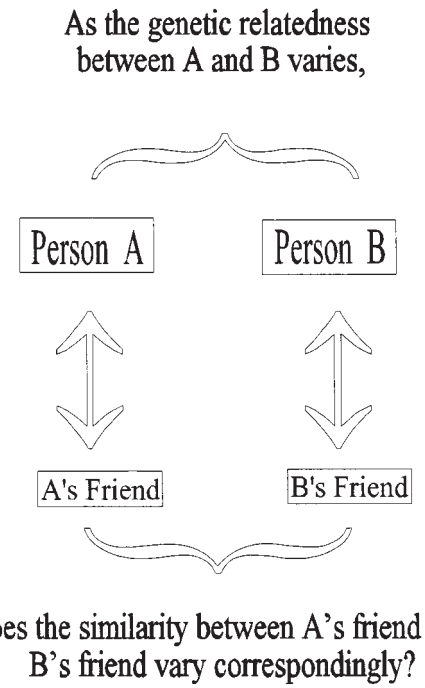

Fig. 1. Research design for studying genetics basis for friendship selection.

ilarity between A's friend and B's friend cannot be attributed to genetic factors alone and because the similarity between A and B also share certain environments.

In this application, we test the hypothesis using data from the Add Health Study. The Add Health Study has taken measures not only on a large number of genetically related individuals but also on their friends themselves. We estimate the correlation between A's friend and B's friend by type of genetic relatedness between A and B after controlling for potential confounding effects. Age can be one such effect. The genetic relatedness within a pair of DZ twins is the same as that within a pair of full siblings. The two types of pairs, however, differ in the age difference between the two members of a pair. This possible age effect can be easily incorporated in the analysis.

We use the following model for the analysis outlined:

$$Y_{ij(t)} = \beta_0 + \beta_1 age_{ij} + u_{j(t)} + e_{ij(t)} \qquad (12)$$

Model (12) is a special case of model (4) presented earlier. Adjusting for age, the correlation between A's friend and B's friend with respect to Y_{ij} for a particular genetic relatedness t is estimated by $\rho_t = \sigma^2_{ut} / (\sigma^2_{ut} = \sigma^2_{et})$. Our hypothesis is that ρ_m is considerably larger than ρ_d and ρ_f, that ρ_d and ρ_f are approximately equal, that ρ_d and ρ_f are larger than ρ_h, and that ρ_h is larger than ρ_c.

In this particular application, the outcome variable Y_{ij} is GPA. We expect that our hypothesis will find more support in data consisting of only same-gender friendships than in data consisting of only different-gender friendships. We have a same-gender friendship when A and A's friend or B and B's friend are of the same gender. When an adolescent looks for a different-gender friend, his or her criterion may be different from those when the adolescent looks for a same-gender friend. Similar GPA may be more important for a same-gender friendship, whereas physical attractiveness may be more important for a different-gender friendship.

Table III presents the estimates of correlations between A's friend and B's friend with respect to GPA by genetic relatedness and by whether the friendship is between the same gender or different gender. The findings are consistent with our expectations. When the friendships are between the same gender, the correlation between MZ twins' friends (.566) is much larger than those between DZ twins' friends (.223) and full siblings' friends (.266), which, in turn, are larger than those for half siblings' friends (.060) and cousins' friends (0). The same pattern is not evident in the different-gender data.

Table III. The Mixed or Multilevel Model Estimates of Correlations Between A's Friend and B's Friend with Respect to Grade Point Average by Genetic Relatedness and by Whether the Friendship Is Between the Same Gender or Different Gender (also see Fig. 1)

	Correlation coefficient	
Genetic relatedness	Same gender	Different gender
MZ twins	.567	.399
DZ twins	.223	.316
Full siblings	.266	.180
Half Siblings	.060	.200
Cousins	.000	.023
Sample size	1924	1386

Application III: Interactions Between Environment and Heritability. It is hypothesized that social conditions modulate the expression of biological or genetic predispositions (Bouchard *et al.*, 1990; Rowe *et al.*, forthcoming; Udry, 1996). Genes provide the potential, but whether the potential can be realized is determined by environmental conditions. The interactions between genetics and environment can be illustrated dramatically by comparing Asian immigrants to the United States and U.S.-born Asians. Studies have found that because of lifestyle and diet, U.S.-born Asians are twice as likely as Asian immigrants to suffer from prostate cancer (Cook *et al.*, 1999) and that Asian American adolescents born in the United States are more than twice as likely to be obese as are adolescents who immigrated recently (Popkin and Udry, 1998).

Within a society, individuals may enjoy different levels of opportunities or face different levels of societal constraint with respect to a particular behavior. It is probably less likely for a minority child growing up in an urban ghetto than a middle-class child of equal genetic potential growing up in an affluent suburban area to succeed in getting a first-rate education and a middle-class income.

In application III, we test the hypothesis that disadvantaged environments decrease the influences of heritability and increase the influences of shared environments on children's intellectual development. This hypothesis can be tested explicitly by comparing the level of heritability in advantaged environments with that in disadvantaged environments via (7) and (9), examining environmental factors one at a time while controlling for the influences of other environmental factors.

Table IV presents fixed effects and the variances of random effects from a mixed or multilevel model of intellectual development measured by PPVT on

Table IV. Fixed Effects and the Variances of Random Effects from a Mixed or Multilevel Model of Intellectual Development on Selected Environmental Variables: MZ Twins, DZ Twins, Full Siblings, Half Siblings, and Cousins

	Random effects (variance components)	Standard errors	Fixed effects (coefficients)	Standard errors
Intercept			89.81	1.48
Income			1.52	0.39
Black	22.22	9.35	−9.87	0.65
Mother education	0.49	0.39	2.46	0.23
AFDC	24.63	17.19	−2.77	1.07
$\sigma^2_{u(m)}$ (MZ twins)	77.25	15.08		
$\sigma^2_{u(d)}$ (DZ twins)	39.84	11.21		
$\sigma^2_{u(f)}$ (full sibs)	25.19	7.68		
$\sigma^2_{u(h)}$ (half sibs)	19.97	12.36		
$\sigma^2_{u(c)}$ (cousins)	5.20	19.75		
$\sigma^2_{e(m)}$ (MZ twins)	45.14	5.71		
$\sigma^2_{e(d)}$ (DZ twins)	81.50	7.10		
$\sigma^2_{e(f)}$ (full sibs)	85.39	5.11		
$\sigma^2_{e(h)}$ (half sibs)	101.73	10.73		
$\sigma^2_{e(c)}$ (cousins)	107.71	19.06		
logL	19427.16			
Sample size	2545			

selected environmental variables using a genetic sample of MZ twins, DZ twins, full siblings, half siblings, and cousins. The mixed model has included four environmental variables: logarithm of family income, black with white as the reference group, mother's education,[1] and Aid to Families with Dependent Children (AFDC). AFDC is a dummy variable coded as 1 if the family received AFDC at the time of the survey and 0 otherwise. The fixed effects of all of these environmental variables are consistent with previous research. Of the variances of the random coefficients of the environmental variables, only black is statistically significant. This random coefficient is interpreted by the calculations presented in Table V.

Using the parameter estimates from Table IV and Eq. (9), we calculated heritability (h^2) and proportions of the variance owing to shared environmental influences among MZ and DZ twins ($c^2_{(md)}$), full siblings ($c^2_{(f)}$), half siblings ($c^2_{(h)}$), and cousins ($c^2_{(c)}$) by race (Table V). The values of the environmental variables other than black are set at the sample means. The heritability among African Americans (.488) with respect to intellectual development is about 14.8% less than that for whites (.573). On the other hand, the shared environmental effects for African Americans are considerably larger than those for whites regardless of the type of clusters. The shared environmental effects for whites are small, at most a few percentage points of the total effects. Assuming that the environments for intellectual development for African Americans are still less favorable than those for whites after controlling for family income, mother's education, and welfare receipt, these results are consistent with our hypothesis. The disadvantaged environments appear to have prevented African Americans from fully realizing their genetic potentials for intellectual development and to have increased the relative importance of shared environmental influences.

Table V. Estimated Heritability (h^2) and Proportions of the Variance Owing to Shared Environmental Influences Among MZ and DZ Twins (c^2_{dm}), Full Siblings (c^2_f), Half Siblings (c^2_h), and Cousins (c^2_c) by Race

	h^2	c^2_{dm}	c^2_f	c^2_h	c^2_c
Blacks*	.488	.214	.144	.243	.18
Whites	.573	.077	−.015	.065	.028

*Other observed variables are set at the sample means.

[1] Mother's education is coded from 0 through 6 and treated as a continuous variable: never went to school = 0, <8th grade = 1, 8–11th grades = 2, high school graduate or equivalent = 3, some college = 4, college graduate = 5, and professional training beyond four-year college = 6.

MODELS FOR COMPLEX FAMILY BEHAVIORAL GENETIC DATA

The Model

Earlier, we show how clustered data vis-à-vis paired data can be handled readily by the mixed or multilevel model. In this section, we propose a version of the mixed or multilevel model that handles more complex family behavioral genetic data. The complexity arises when three or more siblings in a family form pairs that have different degrees of genetic relatedness. Suppose that a family has siblings A, B, and C. Siblings A and B are a pair of MZ twins. Siblings A and C are a pair of full siblings. So are siblings B and C. Clearly, this family of siblings is no longer clustered data where all possible pairs are genetically related the same way. A more complicated sibling structure in a family may consist of more different sibling pairs or clusters. For example, a family may consist of siblings A, B, C, D, E, F, and G. A and B are a pair of MZ twins. C and D are a pair of DZ twins. E and F are full siblings to each other and to A, B, C, or D. G is a half sibling to everybody else.

Our model for the complex sibship structure in a family is another special case of model (1). We start with a model equivalent to model (4) that estimates both the within and between variances for the five different types of siblings while adjusting for the environmental influences:

$$Y_{ij(t)} = \beta_0 + \beta_1 x_{1ij} + \ldots + \beta_P x_{Pij} + {}^*u_{j(c)} + {}^*u_{j(h)} + {}^*u_{j(f)} + {}^*u_{j(d)} + {}^*u_{j(m)} + e_{ij(t)} \quad (13)$$

where $\mathrm{Var}({}^*u_{j(c)}) = {}^*\sigma^2_{u0(c)}$, $\mathrm{Var}({}^*u_{j(h)}) = {}^*\sigma^2_{u0(h)}$, $\mathrm{Var}({}^*u_{j(f)}) = {}^*\sigma^2_{u0(f)}$, $\mathrm{Var}({}^*u_{j(d)}) = {}^*\sigma^2_{u0(d)}$, and $\mathrm{Var}({}^*u_{j(m)}) = {}^*\sigma^2_{u0(m)}$. The SAS codes for model (13) can be found in Appendix 3. The model assumes $\rho_c < \rho_h < \rho_f < \rho_d < \rho_m$. Model (13) can also be used to analyze clustered data. The relationship between model (4) and model (13) can be shown through the relationship of the between variances of model (4) and model (13) when the two models are used to analyze the same data:

$$\begin{aligned}
\sigma^2_{u0(c)} &= {}^*\sigma^2_{u0(c)} \\
\sigma^2_{u0(h)} &= {}^*\sigma^2_{u0(c)} + {}^*\sigma^2_{u0(h)} \\
\sigma^2_{u0(f)} &= {}^*\sigma^2_{u0(c)} + {}^*\sigma^2_{u0(h)} + {}^*\sigma^2_{u0(f)} \\
\sigma^2_{u0(c)} &= {}^*\sigma^2_{u0(c)} + {}^*\sigma^2_{u0(h)} + {}^*\sigma^2_{u0(f)} + {}^*\sigma^2_{u0(d)} \\
\sigma^2_{u0(m)} &= {}^*\sigma^2_{u0(c)} + {}^*\sigma^2_{u0(h)} + {}^*\sigma^2_{u0(f)} + {}^*\sigma^2_{u0(d)} + {}^*\sigma^2_{u0(m)}
\end{aligned} \quad (14)$$

In this section, we only present matrices for model (13) that differ from those for model (4). To accommodate the complex family sibship structures, we construct a new design matrix $\boldsymbol{Z}$ that has z_js on the main diagonal and 0s elsewhere. The z_j is family specific. For a family that has one cousin, one half sibling, one full sibling, two DZ twins, and two MZ twins, the z_j is

$$\boldsymbol{z}_j = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

where the five columns are the five indicator variables z_c, z_h, z_f, z_d and z_m, representing the coding for the cousin, the half sibling, the full sibling, the two DZ twins, and the two MZ twins, respectively; the rows within a family are sorted so that genetically less-related siblings go first; alternatively, the rows can be sorted so that the genetically most-related siblings go first:

$\boldsymbol{u} = ({}^*u_{1(c)}, {}^*u_{1(h)}, {}^*u_{1(f)}, {}^*u_{1(d)}, {}^*u_{1(m)}, \ldots, {}^*u_{N(c)}, \ldots {}^*u_{N(m)})'$;

and

$$G = I_N \otimes \begin{pmatrix} {}^*\sigma^2_{u0(c)} & & & & \\ & {}^*\sigma^2_{u0(h)} & & & \\ & & 0 & & \\ & & & & {}^*\sigma^2_{u0(m)} \end{pmatrix}$$

To include the random effects of environmental influences, we extend model (13) to obtain

$$Y_{ij(t)} = \beta_0 + \Sigma_{p=1}^{p} \beta_p x_{pij} + \Sigma_{p=1}^{p} u_{pj} x_{pij} + \Sigma_{t=c}^{m} {}^*u_{j(t)} + e_{ij(t)} \quad (15)$$

The SAS codes for this model are given in Appendix 4. Similar to model (7), model (15) does not allow the random slopes to vary by sibling types, although varying these slopes is possible. For a family that has one cousin, one half sibling, one full sibling, two DZ twins, and two MZ twins, the z_j for (15) is

$$z_j = \begin{pmatrix} x_{11j} & \ldots & x_{P1j} & 1 & 0 & 0 & 0 & 0 \\ x_{12j} & \ldots & x_{P2j} & 1 & 1 & 0 & 0 & 0 \\ x_{13j} & \ldots & x_{P3j} & 1 & 1 & 1 & 0 & 0 \\ x_{14j} & \ldots & x_{P4j} & 1 & 1 & 1 & 1 & 0 \\ x_{15j} & \ldots & x_{P5j} & 1 & 1 & 1 & 1 & 0 \\ x_{16j} & \ldots & x_{P6j} & 1 & 1 & 1 & 1 & 1 \\ x_{17j} & \ldots & x_{P7j} & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

and the matrices $\boldsymbol{u}$ and $\boldsymbol{G}$ are, respectively,

$$\boldsymbol{u} = (u_{11}, \ldots, u_{1p}, {}^*u_{1(c)}, {}^*u_{1(h)}, {}^*u_{1(f)}, {}^*u_{1(d)}, {}^*u_{1(m)}, \ldots, {}^*u_{N1}, \ldots {}^*u_{NP}, {}^*u_{N(c)}, \ldots, {}^*u_{N(m)})'; \quad \text{and}$$

$$G = I_N \otimes \begin{pmatrix} \sigma^2_{u1} & & & & & & \\ & 0 & & & & & \\ & & \sigma^2_{uP} & & & & \\ & & & {}^*\sigma^2_{u0(c)} & & & \\ & & & & 0 & & \\ & & & & & {}^*\sigma^2_{u0(m)} & \end{pmatrix}$$

Using parameter estimates from (15), we can calculate the correlation between two siblings of the same type:

$$\rho_{(t)}(Y_{ij}, Y_{i'j} \mid x_j) = \frac{\sum_{t=c}^{m} {}^*\sigma^2_{u0(t)} + \sum_{p=1}^{P} \sigma^2_{up} x^2_{pj}}{\sigma^2_{e0(t)} + \sum_{t=c}^{m} {}^*\sigma^2_{u0(t)} + \sum_{p=1}^{p} \sigma^2_{up} x^2_{pj}} \qquad (16)$$

Models (13) and (15) may be conceptualized as two-level multilevel models with siblings as level-1 units and families as level-2 units. However, these are not conventional multilevel models, in which all pairs of observations in the same cluster are correlated in the same way. In contrast, models (13) and (15) allow the observations in the same family to have various degrees of correlation. Our models for complex family behavioral genetic data may be considered an extension of conventional multilevel models.

Applications of the Models

In this section, we apply models (13) and (15) to the complex family data from the Add Health Study. This sample includes not only white and black students but also students from other ethnic groups. Of the 4999 students in the sample, more than 13% are found to live in a family that has a complex sibling structure or a family that consists of more than one type of sibling. The actual percentage of the students living in a family with a complex sibling structure must be substantially higher than 13%. Because of the study design, the Add Health data underrepresent the complexity of sibship structure in American families. The Add Health Study is a school- rather than household-based study. The study includes students from the 7th grade to the 12th grade. Siblings from the same family were included in the study only when they were in these grades and when they went to the schools under study at the time of the survey. Our methodology should be considerably more useful for sibling data from household surveys.

Table VI. Estimated Fixed Effects and Variances of Random Effects and Their Standard Errors from Two Mixed or Multilevel Models of Intellectual Development Using Complex Family Behavioral Genetic Data. The Model Numbers Correspond to the Equation Numbers in the Text. Model (13) Is a Basic Variance Decomposition Model and Model (15) Estimates the Interaction Between Environment and Heritability

	Model (13)	Model (15)
Fixed effects		
Intercept	98.4 (.24)	102.3 (.28)
Minorities		−9.22 (.45)
Random effects		
${}^*\sigma^2_{u0(m)}$ (MZ twins)	0.0	0.0
${}^*\sigma^2_{u0(d)}$ (DZ twins)	54.0 (17.8)	71.0 (16.4)
${}^*\sigma^2_{u0(f)}$ (full sibs)	24.5 (13.3)	32.5 (11.1)
${}^*\sigma^2_{u0(h)}$ (half sibs)	0.0	0.0
${}^*\sigma^2_{u0(c)}$ (cousins)	31.1 (16.1)	28.3 (13.3)
${}^*\sigma^2_{u0(u)}$ (unrelated)	59.3 (11.2)	14.7 (9.3)
$\sigma^2_{e(m)}$ (MZ twins)	48.9 (4.2)	49.2 (4.3)
$\sigma^2_{e(d)}$ (DZ twins)	101.7 (6.4)	100.9 (6.3)
$\sigma^2_{e(f)}$ (full sibs)	93.2 (4.1)	95.7 (4.3)
$\sigma^2_{e(h)}$ (half sibs)	117.8 (9.3)	118.2 (9.3)
$\sigma^2_{e(c)}$ (cousins)	161.4 (19.9)	165.8 (20.1)
$\sigma^2_{e(u)}$ (cousins)	127.6 (10.4)	132.1 (11.0)
Minorities		35.1 (7.6)
−2logL	40058.2	39657.3
Sample size	4999	4999

In Table VI, we present the estimated fixed effects and variances of random effects and their standard errors from two mixed or multilevel models of intellectual development. Model (13) is a basic model of variance decomposition. Model (15) estimates parameters that may be used to calculate the differential heritability associated with the white population and the nonwhite population. By the likelihood ratio test, the two additional parameters of nonwhite and the variance of nonwhite are highly significant. Table VII presents correlation

Table VII. Correlation Coefficients Calculated from Parameters Estimated from Models (13) and (15) Using Complex Family Behavior Genetic Data

	Correlation coefficient		
		Model (15)	
Type of genetic relatedness	Model (13)	Majorities	Minorities
MZ twins	.775	.748	.786
DZ twins	.624	.590	.643
Full siblings	.550	.441	.536
Half siblings	.431	.266	.397
Cousins	.360	.205	.320
Unrelated	.317	.100	.274

coefficients calculated from parameters estimated from models (13) and (15). Model (15) yields two sets of correlation coefficients, one for the white population and the other for the nonwhite population. The implied heritabilities for the white and nonwhite populations are .32 and .28, respectively, with that for the white population only slightly larger than that for nonwhite population. This may be explained partially by the large number of Asians in the nonwhite population.

COMPARING WITH STRUCTURAL EQUATION MODELS

In this section, we compare results from structural equation models, mixed or multilevel models, and Pearson's correlation analysis. Structural equation models (Neale and Cordon, 1992) have been established as the main methodological approach for behavioral genetic analysis. For the comparative analysis, we have extracted a sample of MZ (179) and DZ (353) twin pairs from the Add Health Study.

The basic structural equation model for paired data (figure omitted) has three latent variables representing, respectively, genetic, shared environmental, and nonshared environmental influences. The model usually has two outcome variables for the two siblings in a pair. Three parameters would be estimated for the effects of the latent variables on the outcome variables. Two additional parameters would be used to reflect the genetic relatedness and the common environment, respectively, of the two siblings in a pair. A SEM variance decomposition model would include two setups like this, one for MZ twins and the other for DZ twins. A SEM heritability/environment interaction model with the environment indexed by two-category dummy variables would include four setups with one for each combination of sibling type and environment category.

The first three columns of Table VIII display correlation coefficients from the variance decomposition model from the structural equation analysis, the mixed or multilevel analysis, and Pearson's correlation analysis. The three sets of results are almost identical. The fourth and fifth columns of Table VIII provide correlation coefficients from a heritability/environment interaction model from the structural equation analysis and the mixed or multilevel analysis. The social environment is indexed by a two-category variable on mother's education. L-edu indicates a high school education or low, and H-edu indicates at least some college education. The heritability estimates for the two education groups may be calculated from these correlation estimates. The results from the two types of analysis are very similar.

The SEM heritability/environment interaction model is quite unrestrictive in the sense that it estimates one set of parameters for each group defined by sibling type and education. In comparison, the mixed models we presented in Tables IV and VI are more restrictive, in which the random effects of environmental variables are assumed to be invariant across sibling type. In order to make the two analyses comparable, we estimated a mixed model for Table VIII that allows the random effect for education to vary by sibling type. It should be pointed out that the SEM analysis is not insensitive to the order of the two siblings in a twin pair. A change of the order for some or all of the twin pairs may well change the results substantially.

CONCLUDING REMARKS

Through the three applications, we have demonstrated some of the advantages of the mixed model for behavior genetic studies. The mixed model appears to be a promising alternative to existing approaches usu-

Table VIII. Comparing Results from Structural Equation Models (SEM), Mixed Models, and Pearson's Correlation Analysis of Intellectual Development

	Correlation coefficients				
	Variance decomposition models			G*E interaction models	
Types of siblings	SEM	Mixed	Pearson (pairs)	SEM	Mixed
MZ twins	.75	.76	.76 (179)		
DZ twins	.51	.50	.50 (353)		
MZ twins/H-edu				.77	.75
DZ twins/H-edu				.51	.51
MZ twins/L-edu				.73	.72
DZ twins/L-edu				.47	.46

ally employed in behavior genetic studies. Various versions of the mixed model can be estimated by SAS, HLM, Mln, and other commercial statistical packages. An important issue that needs to be dealt with is specific hypothesis testing regarding h^2 and c^2, heritability and shared environmental influences, respectively. Rather than estimated directly by the mixed model, these two quantities are calculated on the basis of several model-estimated parameters. Conventional straightforward hypothesis testing, therefore, is unavailable. Our preliminary work has shown that the delta method and bootstrap resampling are two feasible approaches to this issue, but both require additional computing and programming. The applications described in this article have probably only touched a small portion of what the mixed model is capable of in the context of behavior genetic analysis. Our prediction is that the application list will be extended quickly once social scientists interested in behavior genetic analysis become familiar with the methodology.

APPENDIXES

(1) SAS codes for model (4)

```
PROC MIXED NOCLPRINT NOITPRINT COVTEST;
CLASS CLUSTER_ID TYPE;
MODEL PPVT = X1 X2 / SOLUTION;
RANDOM INTERCEPT / SUBJECT = CLUSTER_ID
GROUP = TYPE;
REPEATED / GROUP = TYPE;
RUN;
```

In line one, NOCLPRINT and NOITPRRINT are output controls. With COVTEST, PROC MIXED provides ***z*** test results for the random parameters. Line 2 declares CLUSTER_ID and TYPE as categorical variables. In line 3, we have the outcome variable on the left side of the equal sign and the list of observed variables on the right side whose fixed effects will be estimated. With SOLUTION, SAS prints out the estimates for fixed effects. Line 4 tells SAS to estimate a variance for random effects at the cluster level with each cluster indicated by CLUSTER_ID, and SAS estimates one such variance for each sibling TYPE. Line 5 tells SAS to estimate a variance for the random errors at the individual level, and SAS estimates one such variance for each TYPE of siblings.

(2) SAS codes for model (7)

```
PROC MIXED NOCLPRINT NOITPRINT COVTEST;
CLASS CLUSTER_ID TYPE;
MODEL PPVT = X1 X2 / SOLUTION;
RANDOM INTERCEPT / SUBJECT = CLUSTER_ID
   GROUP = TYPE;
RANDOM X1 X2 / SUBJECT = CLUSTER_ID;
REPEATED / GROUP = TYPE;
```

These SAS codes are the same as the previous ones except the new line 5 that asks SAS to estimate random slopes for X1 and X2 at the cluster level.

(3) SAS codes for model (13)

```
DATA A;
SET IN.SIBDATA;
Z1 = 1;
Z2 = (TYPE > 1);
Z3 = (TYPE > 2);
Z4 = (TYPE > 3);
Z5 =(TYPE = 5);
PROC MIXED NOCLPRINT NOITPRINT COVTEST;
CLASS FAMILY_ID TYPE;
MODEL PPVT = X1 X2 / SOLUTION;
RANDOM Z1 – Z5 / SUBJECT = FAMILY_ID TYPE
   = UN(1);
REPEATED / GROUP = TYPE;
```

Line 2 points to the input SAS dataset. The next five lines creates the Z matrix for model (13). Line 11 asks SAS to estimate the five variances described in model (13). These variances need to be cumulated to obtain the between variance for each TYPE of sibling.

(4) SAS codes for model (15)

```
DATA A;

SET IN.SIBDATA;
Z1 = 1;
Z2 = (TYPE > 1);
Z3 = (TYPE > 2);
Z4 = (TYPE > 3);
Z5 = (TYPE = 5);
PROC MIXED NOCLPRINT NOITPRINT COVTEST;
CLASS FAMILY_ID TYPE;
MODEL PPVT = X1 X2 / SOLUTION;
RANDOM Z1–Z5 / SUBJECT = FAMILY_ID TYPE
   = UN(1);
RANDOM X1 X2 / SUBJECT = FAMILY_ID;
REPEATED / GROUP = TYPE;
```

ACKNOWLEDGMENTS

This work was supported by the Department of Sociology and the Carolina Population Center at the University of North Carolina at Chapel Hill, the W. T.

Grant Foundation, and the American Sociological Association's Fund for the Advancement of the Discipline Award supported by the American Sociological Association and the National Science Foundation.

REFERENCES

Bearman, Peter S., Jones, Jo, and Udry, J. Richard (1997). *The National Longitudinal Study of Adolescent Health: Research Design* [WWW document]. URL: http://www.cpc.unc.edu/projects/addhealth/design.html

Berkowitz, William R. (1969). Perceived height, personality, and friendship choice. *Psychological Reports* **24**:373–374.

Berscheid, Ellen, and Walster, Elaine (1978). *Interpersonal Attraction*, 2nd ed., Random House, New York.

Bouchard, Thomas J., Jr., Lykken, David T., McGue, Matthew, Segal, Nancy L., and Tellegen, Auke (1990). Sources of human psychological differences: The Minnesota Study of Twins Reared Apart. *Science* **250**:223–228.

Bryk, Anthony S., and Raudenbush, Stephen W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods SAGE*, Newbury Park, CA.

Cherny, S. S., DeFries, John C., and Fulker, David W. (1992). Multiple regression analysis of twin data: A model-fitting approach. *Behavior Genetics* **21**:489–497.

Cook, L. S., Goldoft, M., Schwartz, S. M., and Weiss, N. S. (1999). Incidence of adenocarcinoma of the prostate in Asian immigrants to the United States and their descendants. *Journal of Urology* **161**:152–155.

DeFries, John C., and Fulker, David W. (1985). Multiple regression analysis of twin data. *Behavior Genetics* **15**:467–473.

Eaves, L. J., and Gale, J. S. (1974). A method for analyzing the genetic basis of covariation. *Behavior Genetics* **4**:253–267.

Eaves, L. J., Last, K. A., Young, P. A., and Martin, N. G. (1978). Model fitting approaches to the analysis of human behavior. *Heredity* **41**:249–320.

Goldstein, Harvey (1987). *Multilevel Models in Educational and Social Research*, Griffin, London.

Goldstein, Harvey (1995). *Multilevel Statistical Models*, 2nd ed., Arnold, London; Halstead, New York.

Gottesman, I. I. (1991). *Schizophrenia Genesis: the Origins of Madness*, W. H. Freeman, New York.

Grilo, C. M., and Pogue-Geile, Michael F. (1991). The nature of environmental influences on weight and obesity: A behavior genetic analysis. *Psychological Bulletin* **110**:520–537.

Mason, William M., Wong, G. M., and Entwisle, Barbara (1983). Contextual analysis through the multilevel linear model. *Sociological Methodology* **15**:72–103.

Neale, C. Michael, and Cardon, Lon R. (1992). *Methodology for Genetic Studies of Twins and Families*, Kluwer Academic Publishers, Boston/London.

Plomin, Robert, Owen, Michael J., and McGuffin, Peter (1994). The genetic basis of complex human behaviors. *Science* **264**: 1733–1739.

Popkin B., and Udry, J. R. (1998). Adolescent obesity increases significantly in second and third generation U.S. immigrants: The National Study of Adolescent Health. *Journal of Nutrition* **128**:701–706.

Rodgers, Joseph L., and McGue, Matt (1994). A simple algebraic demonstration of the validity of DeFries-Fulker analysis in unselected samples with multiple kinship levels. *Behavior Genetics* **24**:259–262.

Rowe, David C. (1994). *The Limits of Family Influences: Genes, Experience, and Behavior*, The Guilford Press, New York.

Rowe, David C., Jacobson, Kristen C., and van den Oord, Edwin J. C. G. (forthcoming). Genetic and environmental influences on vocabulary IQ: Parental education level as moderator. *Child Development.*

Searle, S. R. (1971). *Linear Models*, Wiley, New York.

Searle, S. R., Casella, George, and McCulloch, Charles E. (1992). *Variance Components*, Wiley, New York.

Tsuang, Ming T., and Faraone, Stephen V. (1990). *The Genetics of Mood Disorders*, Johns Hopkins University Press, Baltimore.

Udry, J. Richard (1996). Biosocial models of low fertility societies. *Population and Development Review* **22** (Suppl.): 325–336.

Edited by Stacey Cherny