Behavior Genetics, Vol. 35, No. 5, September 2005 (© 2005)
DOI: 10.1007/s10519-005-2868-1

Mixed-Effects Variance Components Models for Biometric Family Analyses

John J. McArdle[2] and Carol A. Prescott[1]

Received 1 Sep. 2004—Final 19 Jan. 2005

Recent substantive research on biometric analyses of twin and family data has used both a biometric path analysis model (PAM) and a biometric variance components model (VCM). Methodological research on these same topics have suggested benefits of using linear structural equation model algorithms (SEMA) as well as mixed effect multilevel algorithms (MEMA). To better understand the potential similarities and differences among these approaches we first highlight the algebraic equivalence between the standard biometric PAM and the corresponding biometric VCM models for family data. Second, we demonstrate how several SEMA programs based on either the PAM or VCM approach produce equivalent estimates for all phenotypic and biometric parameters. Third, we show how the biometric VCM approach (but not the PAM approach) can be easily programmed using current MEMA programs (e.g., SAS PROC MIXED). We then expand the scope of these different approaches to include measured covariates, observed variable interactions and multiple relatives within each family. MEMA software is compared to SEMA software for programming complex models, including the flexibility of data input, treatment of missing data, inclusion of covariates, and ease of accommodating varying numbers of observations (per family or individual).

KEY WORDS: Multilevel mixed-effect models; multivariate twin and family data analysis; structural equation model algorithms.

MIXED-EFFECTS VARIANCE COMPONENTS MODELS (VCMS) FOR BIOMETRIC FAMILY ANALYSES

One fundamental goal of data analysis in behavioral genetics is to partition observed scores into biometric components of variation. Classical research on this topic demonstrated how the same models can be represented in several different ways. Fisher (1918, 1925) described a least squares method to obtain "factors of correlations" or a "variance components model." At about the same time, Wright (1918, 1921) proposed a "path analysis model" (PAM) as a way to decompose the pattern of influences among a set of correlations. The VCM approach was extended by Cattell (1953, 1960) when he proposed a comprehensive series of equations for family data termed "Multiple Abstract Variance Analysis" (MAVA). In subsequent work, Jinks and Fulker (1970) showed how simple versions of the MAVA models could be fitted using VCM based expected mean squares, and Loehlin (1965, 1978) demonstrated how all of these models could be equivalently and clearly represented using the PAM approach. The overview by Li (1968) also makes this equivalence clear:

> "It has been shown that Fisher, in 1918, clearly developed the concept that a correlation between two distant variables is made up of several factors, each factor corresponding to a step

[1] Department of Psychology, University of Southern California, Los Angeles CA, USA.
[2] To whom correspondence should be addressed at Department of Psychology, University of Southern California, Los Angeles, CA 90089, USA. Tel: 213-740-2292; Fax: 213-746-9082

> from one variable to the next. Furthermore, these factors are precisely the path coefficients defined by Wright for the respective situations. Fisher expressed a number of correlations either as a simple product of a chain of factors or as a sum of products, all expressible in the form of a path diagram developed by Wright. In view of these similarities, we conclude that Fisher and Wright were not too far apart in their basic methodology or thinking, although their mathematical languages are very different. It is unfortunate for geneticists like us with less than adequate mathematical training that Fisher did not bother to draw some small diagrams to elucidate his formulas. . .'' (Li, 1968, p. 482).

Contemporary computer programs for *maximum likelihood estimation* (MLE) were initially developed for VCM by Bock and Bargmann (1966), Eaves and Gale (1974) and Eaves *et al.* (1978) and for latent PAM by Martin and Eaves (1977). In subsequent work, McArdle *et al.* (McArdle, 1986; McArdle *et al.*, 1980; McArdle and Goldsmith, 1990) showed how the available *structural equation modeling* (SEM) programs (e.g., LISREL) could be used to carry out BG analyses of the PAM or VCM type even though they were not intended for this purpose. This LISREL approach became very popular (e.g. Boomsma and Molenaar, 1987; Neale and Cardon, 1992), but the Mx program (Neale *et al.*, 2002) has been widely used for complex BG analyses, and the recent Mplus program (Muthén and Muthén, 2004) was shown to be useful for advanced BG applications (Prescott, 2004). In almost all cases, these models are fitted using a SEMA-based mean and covariance structure approach, and the PAM approach is often emphasized.

A great deal of the recent statistical literature uses *mixed-effects multi-level* (MEMA) models (e.g., Bryk and Raudenbush, 1992; Goldstein, 1995; McArdle and Hamagami, 1996). In these contexts, the MEMA calculation can be seen as a VCM-type approach with the added potential of a restricted structure on the means (i.e., as in repeated measures ANOVA). A recent *Behavior Genetics* article by Guo and Wang (2002) describes some interesting BG-type applications of MEMA, including modeling with multivariable family data (using the PROC MIXED procedure from the SAS Institute, 1999). Although Guo and Wang showed how to create MLE for complex correlations (with associated standard errors and tests of fit), they *did not directly calculate biometric estimates*. Instead, the decomposition of genetic and environmental variance components were conducted externally using a separate least-squares algorithm (p. 40). The article concluded by suggesting, ''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.'' (Guo and Wang, 2002, p. 42).

In this presentation we respond to the key BG issue raised by Guo and Wang, and we attempt to bridge any remaining gaps between the PAM and VCM approaches for traditional BG problems using either SEMA or MEMA software. The key difference highlighted here is that *the PAM approach uses fixed ''design of correlations'' with estimated coefficients whereas the VCM approach uses a fixed ''design of coefficients'' with estimated variances*. Given appropriate adjustment for these differences we demonstrate that the PAM and VCM approaches can yield identical information and inferences. These results also show how MEMA models, using the VCM approach with means can permit simultaneous estimation of the biometric parameters using available software (e.g., SAS MIXED). The general approach presented here offers several alternative ways researchers can choose to use BG models and programs, and different options will be useful for more complex problems.

In the section that follows we present the common biometric PAM model based on organizing relationships among pairs of relatives in terms of latent variables in different groups. Next we present the classical VCM approach and focus on one version of this approach which separates all components into orthogonal common and specific contributions. We show it is easy to estimate this using both standard SEMA programs and software (e.g. Mplus or Mx) and mixed effects models (e.g., SAS MIXED and NLMIXED here). For the purposes of demonstration, we generate several simulation examples from known populations for MZ and DZ twin pairs, and we also and deal with some features of incomplete and binary data. In subsequent sections we extend this model to include multiple variables and multiple relatives. These examples are intended to demonstrate how simple and extended BG multivariate analyses can be practical, efficient

and appropriate using any of several forms of widely available SEMA or MEMA computer software.

THE PAM FOR BIOMETRIC DATA

To define some basic terms and establish notation we first present the PAM that is widely used in BG research. We can write the observed score (Y_n) for any person ($n = 1$ to N) as

$$Y_n = \mu + D_n \text{ or } Y_n = \mu + \sigma d_n, \tag{1}$$

where we separate the mean (μ) and deviation score (D_n), and then rescale the deviation into unit variance ($E\{d,\ d\} = 1$) so the coefficient (σ) represents the standard deviation of the scores. In traditional biometric theory we often extend the model to include different families ($f = 1$ to F) and different individuals ($i = 1$ to I_f) within the family

$$\begin{aligned} Y_{f,i} &= \mu + A_i + S_f + E_i \quad \text{or} \\ Y_{f,i} &= \mu + \sigma_a a_i + \sigma_s s_f + \sigma_e e_i \end{aligned} \tag{2}$$

to represent independent sources of the deviations that are additive genetic (A), non-genetic but shared by family members (S), and non-genetic factors specific to the individual (E). (Note: We use the term S rather than C because we use C to define "common" components in many subsequent models). These deviations can also be written with unobserved scores scaled to unit variance ($E\{a,a\} = E\{s,s\} = E\{e,e\} = 1$) so the coefficients ($\sigma_j$) represent the standard deviation of each component. In the simplest version of biometric theory we assume these components are all uncorrelated. More complex versions of this model include components to represent non-additivity (e.g., dominance deviations), examine correlations among these components (e.g., $|E\{A, E\}| > 0$), and consider interactions among components (e.g., A *by* E).

In most biometric analyses we use the collection of data from family members to provide unique estimates of the unknown parameters of equation (2). We assume the pair of relatives ($i = 1$ and 2) comes from an independent set of families and assume: (1) independent environment deviations E for pairs are uncorrelated, (2) shared family deviations S for pairs are perfectly correlated, and (3) additive genetic deviations for pairs are correlated to the extent of the genetic relationship. These assumptions can be expressed as a SEM written for a pair of persons within the same family as

$$\begin{aligned} Y_{f,1} &= \mu + \sigma_a a_{f1} + \sigma_s s_f + \sigma_e e_{f1} \\ Y_{f,2} &= \mu + \sigma_a a_{f2} + \sigma_s s_f + \sigma_e e_{f2,} \text{ and} \\ E\{a_1, a_2\} &= r_a, \end{aligned} \tag{3}$$

where the genetic correlation is assigned for the group (e.g., for MZ pairs, $r_a = 1$, whereas for DZ pairs, $r_a = ½$).

The standard model for a pair of relatives is drawn as a latent variable path diagram in Figure 1a. Although the general structure of this path diagram is familiar to BG researchers, it is

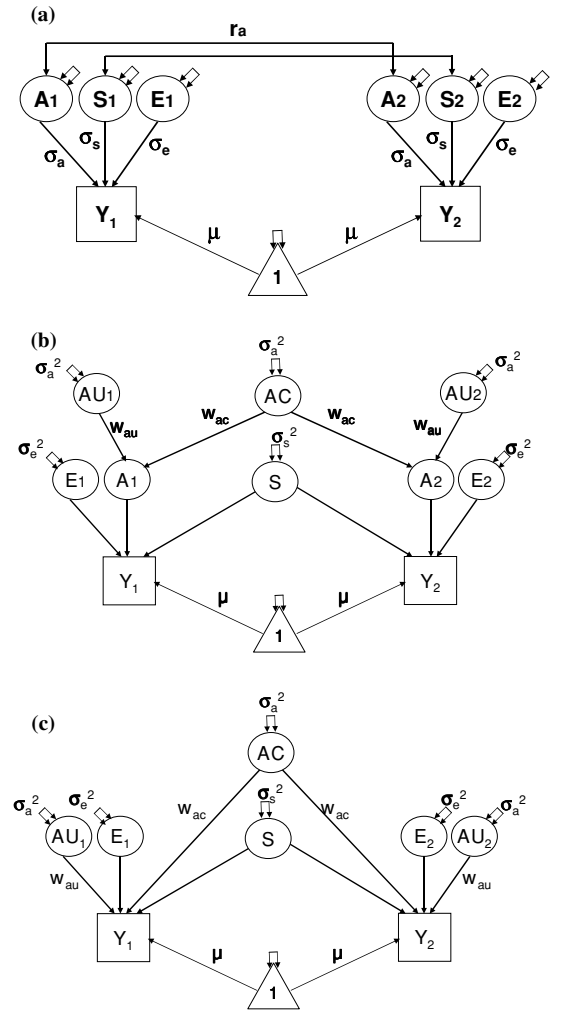

Fig. 1. Alternative specifications for a univariate model for a pair of relatives. All parameters that are not labeled are fixed at 1. (a) Standard SEM specification for twins. $r_{a(mz)} = 1$ and $r_{a(dz)} = 0.5$. (b) SEM representation of an orthogonal variance component specification. Weights are defined by the application (e.g., for twins, $w_{ac(mz)} = 1$, $w_{ac(dz)} = \text{sqrt}(0.5)$, $w_{au(mz)} = 0$, $w_{au(dz)} = \text{sqrt}(0.5)$). (c) SEM reduced-form representation of variance components.

useful to point out that in this particular diagram: (1) All paths that are not explicitly labeled are equal to unity; (2) the variance terms are represented as standard deviations on paths, (3) the means (relations to the constant 1) are included and are equal over all pairs, (4) the variances (two-headed arrows on a variable) are fixed at 1, and (5) the covariances (two-headed arrows connecting two variables) may differ for different groups of relatives.

This representation of pairs of scores within a family generates, in the usual way, equal means and variances for members of each group, but unequal covariances between groups. These structural expectations can be obtained directly from the path diagram in Figure 1a for MZ and DZ groups. It is well known that, by simple substitution of the uncorrelated scores, the equal variances are written as

$$\begin{aligned} \sigma^2_{\mathrm{mz}} &= E\{D_{\mathrm{mz},i}, D_{\mathrm{mz},i}\} = \sigma^2_a + \sigma^2_s + \sigma^2_e \text{ and} \\ \sigma^2_{\mathrm{dz}} &= E\{D_{\mathrm{dz},i}, D_{\mathrm{dz},i}\} = \sigma^2_a + \sigma^2_s + \sigma^2_e, \end{aligned} \tag{4}$$

and the unequal covariances among relatives within each group are written as

$$\begin{aligned} \sigma_{\mathrm{mz},1,2} &= E\{D_{\mathrm{mz},1}, D_{\mathrm{mz},2}\} = \sigma^2_a + \sigma^2_s \text{ and} \\ \sigma_{\mathrm{dz},1,2} &= E\{D_{\mathrm{dz},1}, D_{dz,2}\} = 1/2\sigma^2_a + \sigma^2_s. \end{aligned} \tag{5}$$

This form of the PAM model implicitly restricts the associated variances (σ^2) to be non-negative (i.e., for any value of σ the $\sigma^2 \geq 0$). We return to some implications of this restriction subsequently.

A VCM FOR BIOMETRIC DATA

Some useful features emerge when we restrict our expression of any BG model so "the paths are fixed values for each person." We can rewrite equation (3) for a pair of persons within the same family as

$$\begin{aligned} Y_{f,1} &= \mu + A_{f1} + S_f + E_{f1,} \text{ with} \\ A_{f1} &= w_{\mathrm{ac}} AC_f + w_{\mathrm{au}} AU_{f1}, \text{ and} \\ Y_{f,2} &= \mu + A_{f2} + S_f + E_{f2,} \text{ with} \\ A_{f2} &= w_{\mathrm{ac}} AC_f + w_{\mathrm{au}} AU_{f2}, \end{aligned} \tag{6}$$

where the additive genetic deviation is separated into two deviation scores: (1) AC_f is common for members of the same family, and (2) AU_{fi} is unique to the individual. In this form the weights (W) are fixed at values which indicate the proportion of the additive genetic deviation shared between relatives. This general VCM model is drawn following equation (6) as a nested or higher-order latent variable path diagram in Figure 1b.

A concept that is central to our treatment is how two correlated factors can be re-parameterized as a single shared factor and two unique factors. Figure 1b is useful because it shows the basic theory of the separation of the additive genetic variability for individuals 1 and 2 (A_1 and A_2) into two parts – the part that is *common* to both members of the pair (AC) and the part that is unique to each individual (AU_1 and AU_2). In theory, this separation of additive genetic deviations into common and unique components reflects the result of transmission of parental alleles to offspring. In practice, each of these two new components is assigned the same genetic variance ($\sigma_a{}^2$), we assume these two scores are uncorrelated ($E\{AC,\ AU\} = 0$), and the weights (W) are separated into those which are common to the family (w_{ac}) and specific to the individual (w_{au}). In this formulation, the typical additive genetic correlation (r_{a}) is not included and there are no resulting covariance terms.

It is useful to note that the previous model (6) and Figure 1b can equivalently be written in *reduced form* as

$$\begin{aligned} Y_{f,1} &= \mu + w_{\mathrm{ac}} AC_f + w_{\mathrm{au}} AU_{f1} + S_f + E_{f1,} \text{ and} \\ Y_{f,2} &= \mu + w_{\mathrm{ac}} AC_f + w_{\mathrm{au}} AU_{f2} + S_f + E_{f2,} \end{aligned} \tag{7}$$

and drawn as Figure 1c. Although equations (6) and (7) are formally equivalent, the reduced form is sometime preferable for use by specific computer programs (i.e., SAS MIXED).

In this classic VCM representation, the paths are all fixed weights defined by the application. To insure these weights have no impact on the estimation of the variance terms, they are typically scaled so the sum of squares is unity ($w^2_{\mathrm{ac}} + w^2_{\mathrm{au}} = 1$). As one typical BG example, two individuals in any MZ pair are assumed to share the same genotypes and the same common family environment. This means we can simplify the model for an MZ pair by fixing $w_{ac} = 1$ and 0, and rewrite equation (6) as

$$\begin{aligned} Y_{f,1} &= \mu + A_{f1} + S_f + E_{f1,} \text{ with} \\ A_{f1} &= 1AC_f + 0AU_{f1}, \text{ and} \\ Y_{f,2} &= \mu + A_{f2} + S_f + E_{f2,} \text{ with} \\ A_{f2} &= 1AC_f + 0AU_{f2} \end{aligned} \tag{8}$$

so both the AC and the S latent scores have the identical subscript f, the AU scores are nullified by the zero weights, and the E scores are specific to the individual. In contrast, the assumption of no assortative mating implies that members of DZ pairs share ½ of their genotypes (due to segregation of alleles), and this implication can be represented as fixed weights (of $w_{ac} = \sqrt{1/2}$ and $w_{au} = \sqrt{1/2}$) so the same model for DZ twins is written as

$$\begin{aligned} Y_{f,1} &= \mu + A_{f1} + S_f + E_{f1,} \text{ with} \\ A_{f1} &= \sqrt{1/2}\, AC_f + AU_{f1}, \text{ and} \\ Y_{f,2} &= \mu + A_{f2} + S_f + E_{f2,} \text{ with} \\ A_{f2} &= \sqrt{1/2}\, AC_f + AU_{f2}. \end{aligned} \tag{9}$$

In this model the AC and S scores have subscript f, indicating they are common among all members of the family, but there are now both AU and E scores which are unique to the individual. Notice that the weights for the additive deviations have changed from the PAM version presented in equation (3), but the additive genetic parameters (σ_a^2) are repeated across all AC and AU components, and the partitioning of the scores for the DZ pairs allows the separation of both common and specific genetic components.

An analogous approach could be used to estimate genetic dominance. Two new latent scores, D_i can be added for each individual, decomposed into common DC and unique DU_i, with both assigned equal variance (i.e., σ_d^2). The analysis of MZ–DZ twins would require corresponding weights of 1 and 0 in MZ pairs and weights of $\sqrt{1/4}$ and $\sqrt{3/4}$ and for DC and unique DU_i respectively.

The VCM approach of equation (6) represents all of the parameters of the PAM model of equation (3) and generates the same covariance expectations as in equations (4) and (5). In simple terms, the PC free paths and fixed covariances have been reparameterized by the VCM fixed weights on the paths and free variances. In applications with more complex family configurations, alternative weighting schemes can be used to define orthogonal variance components (described later). It is also useful to note that this VCM approach, in contrast to the more common PAM approach, more closely resembles standard uses of MEMA statistical software, and we turn to these issues next.

PROGRAMMING THE PAM AND VCM BIOMETRIC MODELS

The SEMA software which allow MLE of the mean and variance components with multiple groups can be used to estimate the parameters of both the PAM and VCM approaches (e.g., AMOS, Arbuckle and Wothke, 1999; LISREL, Joreskog and Sorbom, 1999; Mplus, Mx). In general, we do not expect any differences in the results of these SEMA programs for the same model fitted to the same data. Of course, the data setup may vary depending upon the use of various options, such as double-entry of twin scores. Also, the model results may vary slightly depending upon the choice of numerical algorithm used, the convergence criteria, the estimation technique used for standard errors, and any additional boundary constraints on the parameters. To examine these practical issues we present some basic simulation results.

We first defined a population where $\mu = 0$, $\sigma_a^2 = 50$, $\sigma_s^2 = 10$, and $\sigma_e^2 = 40$, and simulated 10 sets of data from $N_{\text{mz}} = 1001$ and $N_{\text{dz}} = 1001$ pairs of twins. It may be useful to note that we used the VCM model to generate the family data. That is, and in general, we do not directly simulate data from a specific pattern of correlations as in equation (3), but these correlations arise from a regression model or a higher-order common factor model such as equation (6). An example of the resulting summary statistics is listed in Table Ia for one simulated dataset. We then fit models to all sets of simulated data using four methods:

1. *SEMA-PAM*: In a first SEMA path analysis we use a fixed set of correlations over two groups to provide estimates of one mean (μ) and three standard deviations (σ_a, σ_s, and σ_e). The estimated components (Table Ib) are consistent with the sample statistics shown in Table Ia. Squaring the parameter estimates produces variance components of $\sigma_a^2 = 4.60$, $\sigma_s^2 = 1.51$, $\sigma_e^2 = 3.84$, with a total score variance of 9.95, and resulting proportions of variance of $\sigma_a^2 = 46\%$, $\sigma_s^2 = 15\%$, and $\sigma_e^2 = 39\%$. The identical estimates were obtained using both the Mplus and Mx programs.

2. *SEMA-VCM*: The second set of programs was written using equation (6) with fixed parameters (weights W) and equality assumptions (σ_a^2) both within pairs (twins 1 and 2) and across the two zygosity groups. The MLE parameters for the complete data are listed in Table Ib. The SEMA-VCM approach adequately recovers all of the parameters

Table I. Summary Statistics for Simulated Twin Data (N = 1001 for each Group)

	MZ Pairs		DZ Pairs	
	Twin 1	Twin 2	Twin 1	Twin 2
(a) Complete data				
N	1001	1001	1001	1001
Means	−0.08	−0.08	−0.14	−0.01
SD	2.99	3.08	3.28	3.22
Correlations	0.592		0.400	
(b) Random drop out				
N	753	774	785	799
Means	−0.07	−0.08	−0.07	−0.06
SD	3.02	3.12	3.31	3.24
Correlations	0.623		0.363	
	(Npr = 583)		(Npr = 620)	
	0.610 (MAR)		0.362 (MAR)	
(c) Genetic selection				
N	760	755	733	752
Means	0.66	0.73	0.84	0.92
SD	2.76	2.81	2.92	2.81
Correlations	0.502		0.293	
	(Npr = 703)		(Npr = 603)	
	0.515 (MAR)		0.297 (MAR)	

MAR estimates are MAR from Mplus output.

of the VCM biometric model in the same way as the previous SEMA-PAM. This is not a surprising result because the simulation model was equivalent to the estimation model.

3. *MEMA-VCM*: A series of mixed effects analyses were carried out using standard mixed effects software (SAS PROC MIXED). In the current version of this program, the MODEL statement and various RANDOM statements can be used to specify alternative models for sources of family resemblance. The basic MEMA-VCM programming devices used here follow the reduced form of equation (7), except that, due to the way SAS PROC MIXED reads the data, they are organized at the individual (rather than family) level. Therefore, each individual has parameters representing each of the specific components. In the case of families of size two, we rewrite the additive genetic score for any person as

$$A_{f,i} = w_a AC_f + w_{u1} AU1_i + w_{u2} AU2_i, \qquad (10)$$

so this component is separated into a genetic score that is common across the family (AC_f) plus two

Table II. Estimated Parameters (MLE) of Univariate Biometric Models using Four Alternative Programs (based on Simulated Data of Table I)

	ML estimates				(z values)			Variance as Percentages		
	μ	σ_a^2	σ_s^2	σ_e^2	z_a	z_a	z_a	σ_a^2	σ_s^2	σ_e^2
(a) Population Values										
	0	5	1	4	(0)	(0)	(0.0)	50.0	10.0	40.0
(b) Complete Cases										
SEMA-PAM	−0.078	2.14^2	1.23^2	1.96^2	(15.0)	(5.8)	(44.0)	46.2	15.2	38.6
SEMA-VCM	−0.078	4.60	1.51	3.84	(7.6)	(2.9)	(22.0)	46.2	15.2	38.6
MIXED-VCM	−0.078	4.60	1.51	3.84	(7.6)	(2.9)	(22.0)	46.2	15.2	38.6
NLMIXED-VCM	−0.078	4.60	1.51	3.84	(7.6)	(2.9)	(22.0)	46.2	15.2	38.6
(c) Random drop out: complete pairs only										
SEMA-PAM	−0.043	2.43^2	0.82^2	1.93^2	(14.0)	(1.9)	(35.0)	57.3	6.5	36.2
VCM programs	−0.043	5.90	0.68	3.73	(7.2)	(1.0)	(17.0)	57.3	6.5	36.2
(d) Random drop out: all available data										
SEMA-PAM	−0.074	2.40^2	0.78^2	1.93^2	(14.0)	(1.8)	(34.0)	57.1	6.1	36.8
VCM programs	−0.074	5.76	0.61	3.71	(7.3)	(0.9)	(17.0)	57.1	6.1	36.8
(e) Genetic selection: complete pairs only										
SEMA-PAM	0.998	1.93^2	0.61^2	1.95^2	(11.0)	(1.2)	(38.0)	47.0	4.8	48.2
VCM programs	0.998	3.71	0.38	3.80	(5.3)	(0.6)	(19.0)	47.0	4.8	48.2
(f) Genetic selection: all available data										
SEMA-PAM	0.727	1.92^2	0.72^2	1.96^2	(11.0)	(1.8)	(38.0)	45.7	6.5	47.7
VCM programs	0.727	3.67	0.52	3.83	(5.3)	(0.9)	(19.0)	45.7	6.5	47.7

z = parameter/standard error.

independent set of additive genetic scores ($AU1_i$ and $AU2_i$) which are unique to an individual in the pair but weighted so equation (10) reduces to (7) (i.e., w_{u2} is 0 for person 1 and w_{u2} is 0 for person 2). The required setup programming is presented in detail in Appendix 1, the PROC MIXED code required is described in detail in Appendix (2), and the numerical results are presented in Table II. The results from MIXED are exactly identical to those obtained with the SEMA-VCM programs or using the traditional SEMA-PAM approach.

4. *NLMIXED-VCM*: An alternative way to run mixed effects models in SAS is to use a non-linear mixed model regression approach available in SAS PROC NLMIXED. The basic model setup is the same as equation (6), the input is extremely flexible and does not require the reduced form (i.e., equation (6) rather than (7)), and the MLE is based on a more general numerical algorithm. The NLMIXED input script is listed in Appendix 4, and the numerical results are identical to the others in Table II. This simply shows that this more flexible program can be used to obtain the same results as all other SEMA and MEMA programs.

Numerical results across programs were also identical for many different values in the simulated datasets. That is, we also compared the parameter estimates obtained using Mplus-VCM and PROC MIXED for several values of $\sigma_a{}^2$ and $\sigma_s{}^2$ (ranging from 0.0 to 0.6). In all cases, all four methods result in identical estimates for the parameter estimates, the standardized proportions of variance (squaring the SEMA-PAM estimates produces the SEMA-VCM estimates), and the likelihood ratio tests of alternative hypothesis (i.e., the A + E, S + E and E Only models). The only notable difference between approaches is that the SEMA-PAM approach applies a lower bound of zero for the variance estimates, whereas this was not made a requirement of the SEMA-VCM or MEMA-VCM approaches (although it is possible).

Differences in estimates between the PAM and VCM approaches are expected only for cases where estimates reach boundaries (e.g., negative variance terms due to model misspecification or zero population values). As pointed out by a reviewer, ''there can be times when these different approaches can result in extremely different results, and these tend to be the cases in which the model has been misspecified. For example, dealing with the four epistastic components (A×A, A×D, D×A and D×D) then adopting a PAM vs. a VCM approach can give very different results if the model is ''mis-specified'' (i.e. there is D×D epistasis, but the model does not allow for it). In some cases, VCMs can therefore be significantly negative, whereas the PAM coefficients would be constrained to be positive.'' These differences can be indexed by a corresponding discrepancy in fit. In these circumstances the VCM approach may be preferred (using either SEMA or MEMA software) because these analyses may permit the researcher better clues about model misspecification.

VARIATIONS ON THE BASIC TWIN MODEL

To investigate the potential for practical differences between SEMA and MEMA programs, we also considered two cases that require more complex numerical evaluations of model likelihoods – (a) Missing Data and (b) Binary Outcomes.

The first comparison of estimation based on raw data can be illustrated by considering results based on two types of ''dropout:'' (1) Dropout that is random with respect to the biometric structure (operationalized as missing values for individuals with low scores; i.e., a z-score ≤ -0.7) where the selection results in only 60% of pairs with complete data, but has little effect on the score statistics (Table Ib). (2) Non-Random dropout data (missing values for individuals with low ''genetic scores'') produced higher score means, reduced variation, and smaller pair correlations than obtained using the original scores (Table Ic). No new computer programs were needed to evaluate the programs because the SEMA and MEMA software used here allow direct entry of raw data with codes for incomplete data. The different models and algorithms produce identical estimates for the analyses based on these more complex datasets (Table IIc–f]). Most of the ''MLE correction'' available from using the incomplete data occurs for the estimates of the variance and means (compare Table IId to IIc and IIb). Including the incomplete pair information from the twins without observed cotwins (Table IIf) produces MLE adjusted estimates of the mean and total variance but still shows the bias expected from this selection process. In principle, these results demonstrate that all the models and algorithms studied here can treat the incomplete data problem in the same way (i.e., MLE based on *Missing at Random* assumptions).

The second type of problem we studied was based on the biometric decomposition of data where

the outcome is a binary variable (0 or 1). Simulated data were generated based on a model where continuous normal deviates were created to follow a standard twin model, but then the continuous outcome was categorized into two classes based on a normal threshold (i.e., $\tau = 1.0$). This was created for several different biometric models and different threshold values (for details, see Prescott, 2004). SEMA computer software to handle this kind of estimation problem has been available for many years, and we used the Mplus program with a Probit estimation (as in Prescott, 2004). There are many MEMA programs that could be used for this problem (e.g., using a GLIMMX macro with MIXED). For simplicity of presentation we used a new VCM variation on the SAS NLMIXED. The additional scripts presented in Appendix 4 include the same biometric decomposition as before (Appendix 3) but add a Logit or Probit score model based on a Bernoulli error distribution (i.e., BINARY; see Powers and Xie, 2000). Our results showed the numerical computations for binary outcomes were rapid using the Mplus-PAM and Mplus-VCM but were decidedly slower using the NLMIXED-VCM scripts. In contrast to the continuous variable cases, the estimates for binary outcomes were not completely equivalent. While either approach should be able to yield accurate biometric results for equivalent models, the exact equality needs to be verified with future research.

INCLUDING MULTIPLE VARIABLES IN VCMS

A new set of problems for the PAM and VCM comparison emerge when we consider the inclusion of another measured variable X as in the path diagrams of Figure 2. Most of these models are easy to fit in the SEMA-PAM approach, but some present an interesting challenge for the representation of the MEMA-VCM approach. The models presented in this section are described but the results are not described (see Appendix 5 scripts).

As the first model, let us consider adding X as a covariate to the equation for Y. In a first biometric model we write

$$Y_{f,i} = \beta_0 + \beta_1 X_i + Z_{f,i} \text{ and } Z_{f,i} = Az_{fi} + Sz_f + Ez_{fi}, \tag{11}$$

where the β are regression coefficients with unobserved residual Z, and the three biometric latent variables (Az, Sz, Ez) are unique to Y given X (i.e., $Az = A|X$). This model is easily fitted in SEMA programs by adding the X variable to the diagram (see Figure 2a). Although the inclusion of variable Z is not needed in the model fitting, it serves to represent the concepts clearly and it gives a natural label and interpretation for the biometric components. It can also be fitted in MEMA programs by adding a variable X only to the fixed part of the model (i.e., the MODEL statement in MIXED).

In an alternative bivariate model we can consider the biometric structure of both Y and X. The basic starting point for this model is written as

$$\begin{aligned} Y_{f,i} &= \mu_y + Ay_{fi} + Sy_f + Ey_{fi} \text{ and } X_{f,i} \\ &= \mu_x + Ax_{fi} + Sx_f + Ex_{fi}. \end{aligned} \tag{12}$$

where each variable has a biometric structure. This model does not yet define the covariance between Y and X, and there are several alternatives that can be used. One popular way is depicted in the path diagram of Figure 2b. This model assumes we rewrite the model for Y to include a regression on the X latent variables as

$$Y_{f,i} = \mu_y + \beta_a Ax_{fi} + \beta_s Sx_f + \beta_e Ex_{fi} + Az_{fi} + Sz_f + Ez_{fi}. \tag{13}$$

In this model, the covariance of Y and X is used in the form of a regression (β) where three additional biometric terms (β_a, β_s, β_e) are used to decompose the Y variance. This model is easily fitted in SEMA programs by adding several other latent variables to the diagram as in Figure 2b (from McArdle *et al.*, 1980, 1998), and the use of standard deviations as path coefficients provides the boundary restrictions of any "Cholesky" decomposition (i.e., no negative variance terms).

The MEMA programs, by current definition, only handle one outcome variable, but in this model there are two (Y and X). In previous MEMA research it has been shown that this type of bivariate model can be programmed in MEMA programs by adding a new set of "dummy weights" (i.e., 0,1) to separate the Y from the X as part of the outcome variable (e.g., see Goldstein, 1995; McArdle *et al.*, 2002). This means that the key problem with the VCM representation of the covariance in regression terms is that it appears to require estimates based on "free paths" (i.e., β_a). As demonstrated earlier, in MEMA models the random

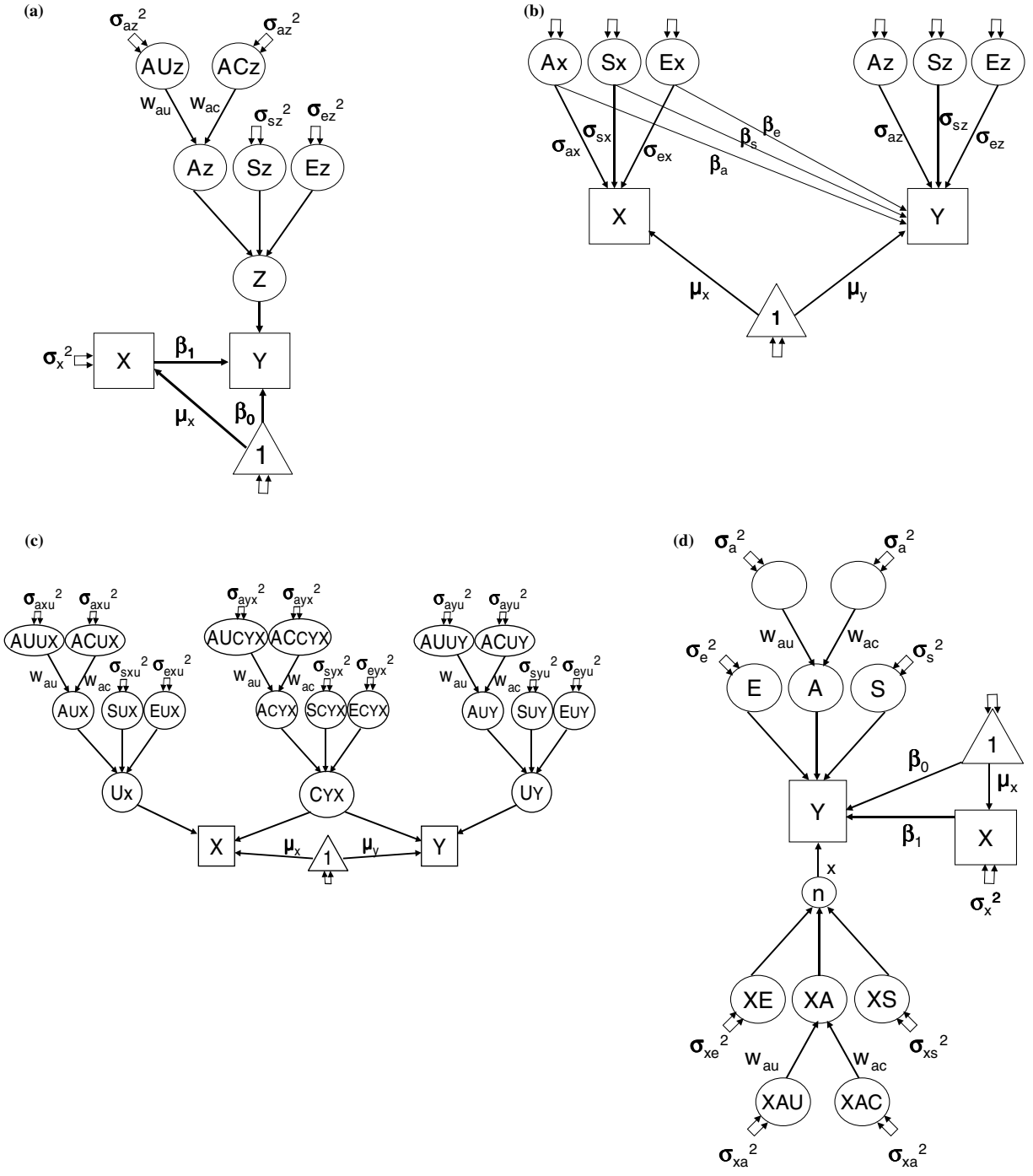

Fig. 2. Alternative models for including a measured variable X. Models are shown for one person. (a) Adding X as a covariate. (b) A popular model for the biometric regression of Y on X. The model shown is the reduced form in which the residual Z is not drawn but is implied and A is not decomposed into AC and AU. (c) A biometric covariance model for Y and X. The biometric covariances are represented as common biometric variances. XU and YU are the unique portions of X and Y after the covariance YXC has been removed. (d) SEMA-VCM representation of an observed variable interaction between X and the biometric components of Y.

variables refer to the fixed weights (or loadings) and the MEMA approach does not (currently) offer a way to introduce random latent variables into the fixed part of the model (i.e., see Appendices 1 and 2).

One new solution to this problem is found by rewriting the biometric model as

$$\begin{aligned} Y_{f,i} &= \mu_y + Cyx_{f,i} + Uy_{f,i;} \text{ and} \\ X_{f,i} &= \mu_x + Cyx_{f,i} + Ux_{f,i;} \text{ with} \\ Cyx_{f,i} &= Acyx_{fi} + Scyx_f + Ecyx_{fi}, \\ Uy_{f,i} &= Auy_{fi} + Suy_f + Euy_{fi,} \text{ and} \\ Ux_{f,i} &= Aux_{fi} + Sux_f + Eux_{fi}. \end{aligned} \tag{14}$$

At a first level, we separate the Y and X scores into common (Cyx) and unique (Uy, Ux) scores, and then, at a second level, we introduce a set of latent variables ($Acyx$, Scy, $Ecyx$) which are common to both Y and X, a set of latent variables (Auy, Suy, Euy) which are unique to Y, and a set (Aux, Sux,-Eux) which are unique to X. As before, if we add a third level of latent variables, we can separate the A terms into AC and AU terms, and then we can then assume all the latent components with variance terms are *orthogonal to one another*. This approach is depicted as a path diagram in Figure 2c, and here it is clear that the biometric decomposition of the covariance of Y and X is represented as a new set of orthogonal variances. Due to this atypical representation of "covariances as variances," we will need to allow these new "variance terms" ($\sigma_{\text{ayx}}{}^2$, $\sigma_{\text{syx}}{}^2$, $\sigma_{\text{eyx}}{}^2$) to take on negative values (as discussed earlier).

In this common variance representation Y is not considered as an outcome of X. Nevertheless, this representation yields the same biometric parameters as the regression approach ($\sigma_{\text{ayx}}{}^2 = \beta_a$), and the total variance of any variable can still be represented in terms of the sum of the common and unique variance (i.e., $\sigma_{\text{ax}}{}^2 = \sigma_{\text{acyx}}{}^2 + \sigma_{\text{aux}}{}^2$). The model of Figure 2c can be fitted using any SEMA program by imposing the appropriate constraints across groups. A script for a SAS MIXED program for this kind of biometric model is presented in Appendix 5 where we explicitly include options so the covariance parameter can be negative.

The need for the addition of an *interaction* into biometric variance models has been considered by many other researchers (Cattell, 1963, 1982; Fulker, 1970; Hendersen, 1975; Mather and Jinks, 1977; Loehlin, 1978; Jinks and Martin *et al.*, 1987; Purcell and Sham, 2002; Eaves and Erkanli, 2003). Some recent research has considered models written in a form similar to

$$Y_{f,i} = \beta_0 + \beta_1 X_{f,i} + A_{fi} + S_f + E_{fi} + X_{f,i}A_{fi}, \tag{15}$$

so the final term represents an "interaction" of observed X and unobserved A. This model is not easy to represent in a standard SEMA path diagram because the XA variable is not strictly latent or manifest. However, an MLE version of this strategy was initially programmed for SEMA using "definition variables" by Neale *et al.* (2002). In a similar way, this kind of variable has been included in a VCM diagram by introducing the "individual measured value as a path" as in Figure 2d here. Although this diagram represents only a restricted form of a general interaction model (i.e., no covariance of A and XA), it is mainly intended to offer a basic framework for further more complex modeling analyses.

In contrast, the traditional MEMA-VCM approach somewhat automatically handles this fixed value by simply using multiplicative weights ($w_n * \mathrm{X}_n$). In the specific case of equation (15) we could simply write

$$\begin{aligned} Y_{f,i} &= \beta_0 + \beta_1 X_{f,i} + A_{fi} + S_f + E_{fi} + X_{f,i}A_{fi}, \text{ with} \\ A_i &= w_{ac}\, AC_f + w_{au}\, AU_{fi}, \text{ and} \\ X_{f,i}A_i &= (X_{f,i}w_{ac})XAC_f + (X_{f,i}w_{au})XAU_{fi}. \end{aligned} \tag{16}$$

In this model there are two additional fixed weights (X^*w_{ac}, X^*w_{au}) whose value depends on both the genetic relatedness of the relative pair and the individual's observed score on X, as well as two additional unobserved deviation scores (XAC, XAU) restricted to have the same variance ($\sigma_{xa}{}^2$). An example of this kind of restricted interaction model is represented in the MEMA -VCM script of Appendix 5. A more general representation of an interaction model is

$$\begin{aligned} Y_{f,i} &= \beta_0 + \beta_1 X_{f,i} + A_{fi} + S_f + E_{fi} \\ &\quad + X_{f,i}(A_{fi} + S_f + E_{fi}), \text{ or} \\ Y_{f,i} &= \beta_0 + \beta_1 X_{f,i} + A_{fi} + S_f + E_{fi} + X_{f,i}A_f \\ &\quad + X_{f,i}S_f + X_{f,i}E_{fi} \text{ or} \\ Y_{f,i} &= \beta_0 + \beta_1 X_{f,i} + A_{fi} + S_f + E_{fi} \\ &\quad + XA_{f,i} + XS_{f,i} + XE_{f,i}, \end{aligned} \tag{17}$$

where three additional latent variables (*XA*, *XS*, *XE*) represent the variation associated with the interaction of observed *X*. The path diagram of Figure 2d shows this model as a higher-order VCM model with a final level of uncorrelated variance components.

One typical application used by twin researchers is a test for sex differences in means and biometric components, and the model of Figure 2d can be used for this purpose. For example, if *X* were coded as 1 for males and 0 for females, the test that $\beta_1 = 0$ is a direct test that males have the same means as females, and the size of the interaction components (σ_{xa}^2, σ_{xs}^2 and σ_{xe}^2) provide tests of sex differences in the biometric components. One advantage of this approach over the scalar sex-limitation PAM model (i.e., that includes an additional source of genetic variance in only one sex) is that the VCM approach does not require the user to decide in advance which sex is expected to have the greater variance. For example, if males have lower genetic variance than females (but were thought to have higher), using the VCM approach (with the coding $X = 1$ if male) will (appropriately) produce a negative value for σ_{xa}^2 in the VCM approach, but merely result in 0 estimates of the male-specific genetic path in the PAM approach. Alternative effect codes (e.g., $X = [1, -1]$ or $X = [\ ½\ ,\ -½\ ]$) can be used to obtain direct estimates of pooled group averages and differences in means and variances.

EXTENSIONS TO INCLUDE MULTIPLE RELATIVES

In the PAM approach, the twin model is typically expanded to include additional relatives by adding sets of latent variables (*A, S, E*) and specifying the expected correlations (or transmission paths) among the latent variables for all the relatives. SEMA programs offer MLE with variable family structures and missing data (e.g., Mx, Mplus), but the code setup time and the numerical estimation time is often prohibitive for large problems. In the VCM approach used here, additional persons can be included in the model by introducing into the data set fixed weights rather than more latent variables. Of course, these extensions also require additional pre-processing of the raw data (beyond Appendix 1), and the required setup work can be extensive. Nevertheless, so the researcher has a chance to compare these approaches, we now illustrate the flexibility of the VCM approach with several examples (see Appendix 6 for SAS MIXED details).

Table IIIa shows the weights needed for the standard two-group twin model (or five-group

Table III. Design Matrices for Programming Twin Models as Variance Component Models

Individual	Genetic variance components						Shared environment			
(a) Standard two-group twin design	*AC*	*AU1*	*AU2*							
MZ1	1	0	0							
MZ2	1	0	0							
DZ1	$\sqrt{1/2}$	$\sqrt{1/2}$	0							
DZ2	$\sqrt{1/2}$	0	$\sqrt{1/2}$							
(b) Adoption design	*AC*	*AU1*	*AU2*	*AU3*	*AU4*	*AU5*				
Bio Sib (1)	$\sqrt{1/2}$	$\sqrt{1/2}$	0	0	0	0				
Bio Sib (2)	$\sqrt{1/2}$	0	$\sqrt{1/2}$	0	0	0				
Bio Sib (3)	$\sqrt{1/2}$	0	0	$\sqrt{1/2}$	0	0				
Adop Sib (4)	0	0	0	0	1	0				
Adop Sib (5)	0	0	0	0	0	1				
(c) *Twin-sibling design*	*AC*	*AM*	*AU1*	*AU2*	*AU3*	*AU4*	*S*	*ST*	*SU1*	*SU2*
MZ Twin 1	$\sqrt{1/2}$	$\sqrt{1/2}$	0	0	0	0	1	1	0	0
MZ Twin 2	$\sqrt{1/2}$	$\sqrt{1/2}$	0	0	0	0	1	1	0	0
DZ Twin 1	$\sqrt{1/2}$	0	$\sqrt{1/2}$	0	0	0	1	1	0	0
DZ Twin 2	$\sqrt{1/2}$	0	0	$\sqrt{1/2}$	0	0	1	1	0	0
Biol Sib1	$\sqrt{1/2}$	0	0	0	$\sqrt{1/2}$	0	1	0	1	0
Biol Sib2	$\sqrt{1/2}$	0	0	0	0	$\sqrt{1/2}$	1	0	0	1

Note: Individual-specific (E) component weights not shown.

without sex differences). These correspond to the weights shown in Figure 1c. The same design matrix and program input could be used to estimate a five-group twin model (if one were not interested in testing for sex effects) and to include triplets who are MZ (but not DZ – see below). A five-group model with sex differences could be estimated by adding common and sex-specific genetic components, or by the method described earlier (Figure 2d).

In principle, adding family members to the model is straight-forward. For each person added there are two necessary changes to the dataset: assigning the appropriate weights for the new family member on the common and unique genetic components and adding weights of zero for the other family members for the new unique component. There is one required change to the MEMA model – adding the new unique genetic component as a random effect. The genetic components which are added are specified to have the same variance as all the other genetic components in the model (σ_a^2). The two basic requirements of the design matrix are: (1) the sum of the squares of the genetic weights for an individual (i.e., all those shown in the same row of Table III) must equal 1, and (2) the correlation of two individuals for each component (i.e., the columns in Table III) is equal to the product of the weights for those individuals.

Table IIIb lists the weights for a study design with biological and/or adopted siblings. The weights for the first two biological siblings are identical to those for DZ twins 1 and 2 in Table IIIa. To add a third biological sibling to this model requires that they be assigned the appropriate weight for the common genetic component ($w_{ac} = \sqrt{1/2}$) and that another weight variable (e.g., w_{au3}) be added for all individuals in the data set to represent the unique genetic variance for the third person. The other relatives in the family would be assigned a 0 value for this weight and the third sibling would be assigned $w_{au3} = \sqrt{1/2}$. Analogous changes could be made to expand the model to include additional biological siblings. Adoptive siblings are added to the model by assigning them $w_{ac} = 0$ (because they are genetically uncorrelated with all other family members) and adding to the dataset another weight variable (e.g., w_{au4}) which would receive a value of 1 for the adoptee and 0 for all other family members.

If siblings labeled 4 and 5 in Table IIIa were related (e.g., they are full siblings of each other, but step-siblings of siblings 1–3), this would be coded by creating a second common genetic factor (e.g., *AC2*) for which siblings 4 and 5 have weights of $\sqrt{1/2}$ and changing their weights for the unique genetic factors (*AU4* and *AU5*) to be $\sqrt{1/2}$. The other family members would be assigned zero weights for all these components.

Applying the correlation principle to Table IIIa shows why the basic model input cannot be used for triplets (except when all three are monozygotic). The problem is more than just the need to add specific *AU* components for the additional family members. For example, if the weights shown were applied to a triplet set which included both MZ and DZ pairings (i.e., two individuals arose from one zygote and the third from another), this weighting would predict that the DZ triplet has a genetic correlation of 0.71 (i.e., 1 * $\sqrt{1/2}$) with his or her co-triplets, rather than the expected value of 0.50. The same problems would arise for a standard twin-sibling design.

Table IIIc is a display of weights which could be used for a twin-sibling design and would accommodate any combination of MZ twins, DZ twins, all triplet types and biological siblings within the same family. Here, all the siblings have a weight of $\sqrt{1/2}$ on the common genetic component (AC) so the expected genetic correlation for the DZ and biological siblings with each other and with their MZ twin siblings is 0.50. However, there is another component (AM) specific to the MZ pair which allows their remaining genetic variance to be correlated and results in the correct expectation of 1.0 for their genetic correlation (see Appendix 6).

The design of the twin-sibling model allows testing for a "special twin environment", where twins have greater environmental similarity than siblings of different ages. A weighting scheme to accomplish this is shown in the far right portion of Table IIIc. All siblings receive a weight of 1 on the *S* component, which has variance σ_s^2. A second component, *ST*, representing the twin environment has weights of 1 for twins and 0 for non-twins, and variance σ_{st}^2. To allow the non-twin siblings to have the same total variance as the twins (as expected), we include residual parameters for each non-twin sibling (SU_1, SU_2) also assigned variance σ_{st}^2, and these are assigned weights of 1 for non-twins and 0 for twins. Unlike the case for the genetic variance components (where $\sigma_{ac}^2 = \sigma_{au}^2$), we do not have an *a priori* hypothesis about the relative magnitude of the *S* and *ST* variances. Thus, the sum of the

shared environment weights in any row are not scaled so that they sum to 1 (as for the genetic components) and we will need to calculate the total shared environment variance as the sum of the *S* and *ST* components. The resulting MLE will represent the shared environmental variance specific to twin siblings.

A variety of other extensions to these sibling models are possible. Additional components could be used to code for other degrees of genetic relatedness (half-siblings, cousins) and other environmental mechanisms such as sibling competition, sex-specific effects, reared-apart twins, and different environments for step-siblings reared together or apart.

DISCUSSION

The main purpose of this paper is to show the complete overlap between the concepts of VCM and PAM biometric models as implemented in SEMA and MEMA computer programs. The SEMA-PAM approach is based on the inclusion of multiple variables with fixed correlations among specific relatives. However, the same parameter estimates can be obtained in SEMA-VCM using fixed weights and orthogonal variance components. This is useful because the available MEMA programs are based solely on the VCM approach. Although the MEMA programs are not always simple to use, they are efficient in both statistical and practical ways, and they have become popular in many areas of behavioral science. These choices between models and algorithms may be important when dealing with more complex multivariate factor analysis, longitudinal applications, and extended family designs, especially analyses dealing with unbalanced and incomplete data (e.g., Loehlin, 1979; Maes *et al*., 1997; McArdle and Prescott, 1996).

In practice, the choice among computer programs will be based on tradeoffs between program features and prior knowledge of the required program coding. Since the evaluation of these tradeoffs was not our main focus, we only presented results from a limited set of readily available software. The SEMA-PAM and SEMA-VCM models presented here were fit to the data using both Mplus and Mx, but we expect other SEMA programs to behave in an identical fashion. In contrast, given the proper weights, the MEMA-VCM approach can be made to be straight-forward and the program input can be simpler than the corresponding SEMA scripts. Other programs for MEMA models can or will soon permit the biometric variance constraints we imposed using SAS MIXED. Programs such as MLM (Goldstein, 1995), HLM (Bryk and Raudenbush, 1992), Splus or R (Pinheiro and Bates, 2000), WinBUGS (Eaves and Erkanli, 2003), and the newer GLLAMM (Skrondal and Rabe-Hesketh, 2004) will all be realistic alternatives for the SAS MIXED code presented here. It is also likely that MEMA packages can or will offer more general estimation functions, so the biometric models with categorical outcomes will also be easy to use (e.g., Prescott, 2004; Skrondal and Rabe-Hesketh, 2004).

The comparison of these techniques can be extended in other ways not discussed here, including the addition of correlations or interactions among the latent components. In practice, unique solutions for these and other biometric models currently depend on the structure and size of the family data available and the resulting limits on biometric inferences (McArdle, 1996). Future work comparing these models and programs may offer insights on how to identify and interpret the parameters of the more complex but appealing biometric models.

APPENDIX: COMPUTER SCRIPTS FOR BIOMETRIC ANALYSIS WITH SAS MIXED

1. SAS script to Create Appropriate Weights for each Pair Type

```
TITLE2 ' Script 1: Building the relational weights needed for the MEVC';
DATA sim_rel; /* assuming pairs are output as separate records */
      SET sim_twin; /* assuming pairs are input on the same record */
      ** MZ twin 1 and 2 weights *;
      IF(mzdzid = 1) THEN DO;
            weightAC=1; weightAU1=0; weightAU2=0;
      END;
      ** DZ twin 1 weights**;
      IF(mzdzid = 0) THEN DO;
                  weightAC=sqrt(.5); weightAU1=sqrt(.5); weightAU2=0;
            END;
      ** assign & output values for twin 1 (MZ and DZ) **;
      y_score = y1_score; person=(famid*10)+1; twinid=1;
      OUTPUT;

      ** DZ twin 2 weights**;
      IF(mzdzid = 0) THEN DO;
            weightAC=sqrt(.5); weightAU1=0; weightAU2=sqrt(.5);
      END;
      ** assign & output values for twin 2 (MZ and DZ) **;
      y_score = y2_score; person=(famid*10)+2; twinid=2;

      ** assign weights for commong factor of E in later models;
          IF (twinid = 1) THEN DO;
                  weightE1 = 1; weightE2 = 0; END;
          IF (twinid = 2) THEN DO;
                  weightE1 = 0; weightE2 = 1; END;
      OUTPUT;

    KEEP person famid mzdzid twinid
            weightAC weightAU1 weightAU2 weightE1 weightE2 y_score;
```

2. SAS MIXED scripts to estimate VCMs

After this pre-processing of the scores (above) we can estimate the full ACE model parameters by running the standard PROC MIXED software using multiple RANDOM statements. An example is given in the following code and explanation:

```
TITLE4 'Script 2a: LEVEL Restricted Covariances to get full A+C+E Twin
Model';
PROC MIXED DATA=sim_rel NOCLPRINT COVTEST METHOD=ML; CLASS famid;
   MODEL y_score =  / SOLUTION DDFM=SATTERTH CHISQ;/* Mean and E variance */

   RANDOM INTERCEPT / SUBJECT=famid TYPE=VC;/* S variance */

      RANDOM weightAC weightAU1 weightAU2 / SUBJECT=famid TYPE=TOEP(1);
 /*A Variance */
      RUN;
```

Step 1: Before running PROC MIXED we assign the weights $W_n = [w_{ac}, w_{au1}, w_{au2}]$: (a) for both MZ members as $W_n = [1, 0, 0]$, (b) for the first DZ member ($n = 1$) as $W = [\sqrt{\frac{1}{2}}, \sqrt{\frac{1}{2}}, 0]$, and (c) for the second DZ member ($n = 2$) as $W = [\sqrt{\frac{1}{2}}, 0, \sqrt{\frac{1}{2}}]$.

Step 2: The individuals' scores within a family are reorganized into a relational database with each record containing (a) family ID (famid = f to F), (b) twin ID (twin = 1 or 2), and (c) the three weights (W).

Step 3: PROC MIXED requires estimation using either METHOD = ML or METHOD = REML, and the CLASS statement must include the family ID (famid = f to F).

Step 4: The MODEL statement must include only the dependent variable (Y_n) (even if multivariate or repeated measures are included; see McArdle and Hamagami, 2003). The statement "MODEL Y = / SOLUTION;" is sufficient in most cases to estimate the mean (μ) and the individual specific environment or error term (σ_e^2), i.e., the E Only Model.

Step 5: The common family variance (σ_S^2) and specific environment (σ_e^2) terms (the CE model) are estimated by adding the statement: "RANDOM INTERCEPT/SUBJECT = famid TYPE = VC;" to the MODEL statement above.

Step 6: A second RANDOM statement is used to estimate the additive genetic variance (σ_a^2). Since the model has fixed weights for the common variance (w_{ac}) and the two individual unique terms (w_{au}) we need to use a device which assures that the variance terms assigned to each weight are "equal and uncorrelated." One way to accomplish this with the current version of PROC MIXED is to write "RANDOM W1 W2 W3/SUBJECT = famid TYPE = TOEP(1);" This particular use of a Toeplitz matrix with a single band results in a single diagonal matrix with equal elements for the additive genetic variance.

Step 7: The full model (A + S + E) is estimated using the MODEL statement and both RANDOM statements (i.e., combining steps 4 and 5), and some of the results are listed below:

```
                    The Mixed Procedure
                    Model Information
Data Set                          WORK.SIM_REL
Dependent Variable                y_score
Covariance Structures             Variance Components,
                                  Banded Toeplitz
Subject Effects                   famid, famid
Estimation Method                 ML
Residual Variance Method          Profile
Fixed Effects SE Method           Model-Based
Degrees of Freedom Method         Between-Within

                      Dimensions
            Covariance Parameters             3
            Columns in X                      1
            Columns in Z Per Subject          4
            Subjects                       2000
            Max Obs Per Subject               2

               Observations Used             4000
               Observations Not Used            0
               Total Observations            4000
                    Iteration History
Iteration    Evaluations       -2 Log Like       Criterion

        0             1     20476.21190162
        1             2     19876.27269444      0.00000002
        2             1     19876.27255286      0.00000000
```

```
                    Convergence criteria met.
                  Covariance Parameter Estimates
                                        Standard         Z
Cov Parm      Subject    Estimate         Error      Value        Pr Z

Intercept     famid        4.9175        0.4966       9.90      <.0001
Variance      famid       0.08766        0.5780       0.15      0.4397
Residual                   4.7832        0.2096      22.82      <.0001

                       Fit Statistics
                  -2 Log Likelihood            19876.3
                  AIC (smaller is better)      19884.3
                  AICC (smaller is better)     19884.3

                       Fit Statistics

                  BIC (smaller is better)      19906.7


                  Solution for Fixed Effects

                              Standard
 Effect       Estimate         Error       DF    t Value    Pr > |t|

        Intercept     0.05920      0.06077     1999       0.97      0.3301
```

Step 8: Alternative models can be fitted by commenting out the input lines or by fixing a parameter at a specific value. For example, a model of additive genetic and individual-specific components (i.e., the A + E model) can be estimated using the MODEL statement and only the second RANDOM statement. It is convenient to simply comment out the needed lines by placing /* before and */ after the statements to be ignored as follows:

```
TITLE4 'Script 2b: Restricted A+E Twin Model by comment';
PROC MIXED DATA=sim_rel NOCLPRINT COVTEST METHOD=ML; CLASS famid;
      MODEL y_score =  / SOLUTION DDFM=SATTERTH CHISQ;
/* Mean and E variance */
/* RANDOM INTERCEPT / SUBJECT=famid TYPE=VC; */ /* No S variance */
      RANDOM weightAC weightAU1 weightAU2 / SUBJECT=famid TYPE=TOEP(1);
/* A Variance */
      RUN;

TITLE4 'Script 2b: Restricted A+E Twin Model by PARM-HOLD';
PROC MIXED DATA=sim_rel NOCLPRINT COVTEST METHOD=ML; CLASS famid;
      MODEL y_score =  / SOLUTION DDFM=SATTERTH CHISQ;
/* Mean and E variance */
      RANDOM INTERCEPT / SUBJECT=famid TYPE=VC;
/* RANDOM INTERCEPT / SUBJECT=famid TYPE=VC; */ /* S variance */
      RANDOM weightAC weightAU1 weightAU2 / SUBJECT=famid TYPE=TOEP(1);
/* A Variance */
      PARMS (5) (0) (5) / HOLD=2; /* fixing Var(S)=0 */ RUN;
```

3. SAS NLMIXED scripts to estimate VCMs

The following NLMIXED code is quite flexible, permitting free-form notation and several levels of higher-order equations. This code easily converged when used on the simulated data and provided confidence intervals for the variance percentages.

```
TITLE4 'Model 3a: LEVEL results with NLMIXED for full A+C+E Twin Model';
PROC NLMIXED DATA=sim_rel;
      yexp = mu + S + weightAC*AC + weightAU1*AU1 + weightAU2*AU2;
      MODEL y_score ~ NORMAL(yexp,ve);
      RANDOM  S AC AU1 AU2 ~ NORMAL([0,0,0,0],
            [vs,
              0, va,
              0, 0, va,
              0, 0, 0, va]) SUBJECT=famid;
      PARMS mu=0 ve=10 va=1 vs=1;
    ESTIMATE 'A% Additive' va / (va + vs + ve);
    ESTIMATE 'S% Shared  ' vs / (va + vs + ve);
    ESTIMATE 'E% Independ' ve / (va + vs + ve);
    RUN;
```

The NLMIXED Procedure

Specifications

Data Set	WORK.SIM_REL
Dependent Variable	y_score
Distribution for Dependent Variable	Normal
Random Effects	S AC AU1 AU2
Distribution for Random Effects	Normal
Subject Variable	famid
Optimization Technique	Dual Quasi-Newton
Integration Method	Adaptive Gaussian Quadrature

Dimensions

Observations Used	4000
Observations Not Used	0
Total Observations	4000
Subjects	2000
Max Obs Per Subject	2
Parameters	4
Quadrature Points	1

Parameters

mu	ve	va	vc	NegLogLike
0	10	1	1	10167.945

Iteration History

Iter	Calls	NegLogLike	Diff	MaxGrad	Slope
1	3	10133.7881	34.15687	125.5491	-278.639
2	4	10005.8629	127.9252	110.0313	−1760.33
3	7	9985.57792	20.28499	108.5373	−622.151
4	8	9952.32506	33.25286	71.24914	−157.23
5	9	9940.90926	11.4158	15.49371	−25.8043
6	11	9939.24957	1.659685	2.011573	−2.62548
7	14	9938.51737	0.732209	2.303175	−0.80018
8	17	9938.19206	0.325306	1.837479	−0.42943

9	19	9938.14145	0.050604	0.674192	−0.09676
10	21	9938.13645	0.005	0.072008	−0.00935
11	23	9938.13628	0.00017	0.016782	−0.00026
12	25	9938.13628	7.96E-6	0.003046	−0.00001

NOTE: GCONV convergence criterion satisfied.

Fit Statistics

-2 Log Likelihood	19876
AIC (smaller is better)	19884
AICC (smaller is better)	19884
BIC (smaller is better)	19907

Parameter Estimates

Parameter	Estimate	Standard Error	DF	*t* Value	Pr > |*t*|	Alpha	Lower
mu	0.05921	0.06077	1996	0.97	0.3300	0.05	-0.05997
ve	4.7832	0.2096	1996	22.82	<.0001	0.05	4.3722
va	0.08785	0.5780	1996	0.15	0.8792	0.05	-1.0457
vs	4.9173	0.4966	1996	9.90	<.0001	0.05	3.9435

Parameter Estimates

Parameter	Upper	Gradient
mu	0.1784	0.003046
ve	5.1943	0.001064
va	1.2214	0.000562
vs	5.8911	-0.00035

Shared Additional Estimates

Label	Estimate	Standard Error	DF	*t* Value	Pr > |*t*|	Alpha	Lower
A% Additive	0.008974	0.05905	1996	0.15	0.8792	0.05	-0.1068
S% Shared	0.5024	0.04723	1996	10.64	<.0001	0.05	0.4097
E% Independ	0.4887	0.02223	1996	21.98	<.0001	0.05	0.4451

Additional Estimates

Label	Upper
A% Additive	0.1248
S% Shared	0.5950
E% Independ	0.5323

4. SAS NLMIXED scripts to estimate binary outcomes with variance components

The following NLMIXED code is used to deal with a binary outcome with twin data. This code converged when used on the simulated data and provided confidence intervals for the variance percentages.

```
TITLE4 'Step 4a: Basic Script with NLMIXED for PROBIT Twin Model';
PROC NLMIXED ;
      yhat = muy + S + weightAC * AC + weightAU1 * AU1 + weightAU2 * AU2;
      phat = PROBNORM(yhat);
      MODEL d50_score ~ BINARY(phat);
      RANDOM  S AC AU1 AU2 ~ NORMAL([0,0,0,0],
            [vs,
              0, va,
              0, 0, va,
              0, 0, 0, va]) SUBJECT=famid;
      PARMS muy=-.03 vs=.2 va=.4;
      * BOUNDS vs > 0, vs < 4, va > 0, va < 4;
      ESTIMATE 'Threshold Probability' PROBNORM(muy);
      ESTIMATE 'Total Variance' (va + vs + 1);
      ESTIMATE 'A% Additive' va / (va + vs + 1);
      ESTIMATE 'S% Shared  ' vs / (va + vs + 1);
      ESTIMATE 'E% Independ' 1 / (va + vs + 1);
    RUN;
```

```
NOTE: GCONV convergence criterion satisfied.
NOTE: PROCEDURE NLMIXED used (Total process time):
      real time           4:28.22
      cpu time            4:16.27
```

The NLMIXED Procedure

Fit Statistics

-2 Log Likelihood	5310.8
AIC (smaller is better)	5316.8
AICC (smaller is better)	5316.8
BIC (smaller is better)	5333.6

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
va	0.7238	0.2244	1996	3.22	0.0013	0.05	0.2836	1.1640	5.291E-6
vc	0.5635	0.1682	1996	3.35	0.0008	0.05	0.2336	0.8934	0.000029
muy	0.05591	0.03453	1996	1.62	0.1056	0.05	-0.01181	0.1236	0.000291

Additional Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
A% Additive	0.3164	0.08446	1996	3.75	0.0002	0.05	0.1508	0.4821
C% Common	0.2464	0.07385	1996	3.34	0.0009	0.05	0.1015	0.3912
E% Independ	0.4372	0.03201	1996	13.66	<.0001	0.05	0.3744	0.5000

```
TITLE4 'Script 4b: Basic Setup with NLMIXED for LOGIT A+C+E Twin Model';
PROC NLMIXED;
      yhat = muy + C + weightAC*AC + weightAU1*AU1 + weightAU2*AU2;
      p = 1 /(1+EXP(-yhat));
      pi=3.141592654; ve = ((pi**2)/3);
      vtot= va + vc + ve;
      MODEL d50_score ~ BINARY(p);
      RANDOM  C AC AU1 AU2 ~ NORMAL([0,0,0,0],
            [vc,
              0, va,
              0, 0, va,
              0, 0, 0, va]) SUBJECT=famid;
      PARMS va=.3 vc=.2  muy=0 ;
    ESTIMATE 'A% Additive' va / vtot ;
    ESTIMATE 'C% Common  ' vc / vtot ;
    ESTIMATE 'E% Independ' ve / vtot ;
    RUN;
```

5. SAS MIXED scripts for the Bivariate-Biometric models

The following MIXED codes are quite flexible, and the uses of this code are described in the text. This code was also used on the simulated data.

```
TITLE4 'Script 5a: Adding a covariate to a full A+S+E Twin Model';
PROC MIXED DATA=sim_rel NOCLPRINT COVTEST METHOD=ML; CLASS famid;
      MODEL y_score = x_score / SOLUTION DDFM=SATTERTH CHISQ;
/* Mean and E variance */
      RANDOM INTERCEPT / SUBJECT=famid TYPE=VC;
/* S variance */
      RANDOM weightAC weightAU1 weightAU2 / SUBJECT=famid TYPE=TOEP(1);
/* A Variance */
      RUN;


TITLE4 'Script 5b: Adding a covariance to a full A+S+E for X and Y';
DATA outdata;
      SET indata;
      #1 1 0 y_score
      #2 0 1 x_score
      RUN;
DATA yx_data;
      INFILE outdata;
      INPUT dy dx yx_score;
      RUN;

PROC MIXED DATA=yx_data NOCLPRINT COVTEST METHOD=ML; CLASS famid twinid;
      MODEL yx_score = dy dx / NOINT SOLUTION DDFM=SATTERTH CHISQ;
/* Mean and E covariance as Ve2 */
      RANDOM dy dx / SUBJECT=famid(twinid) TYPE=VC;
/* S variances for Y and X */
RANDOM INTERCEPT dy dx / SUBJECT=famid TYPE=VC;
/* S variances */
      RANDOM weightAC weightAU1 weightAU2 / SUBJECT=famid TYPE=TOEP(1); /*
Common A Covariance */
      RANDOM weightAC*dy weightAU1*dy weightAU2*dy / SUBJECT=famid
TYPE=TOEP(1); /* A Variance for Y */
      RANDOM weightAC*dx weightAU1*dx weightAU2*dx / SUBJECT=famid
TYPE=TOEP(1); /* A Variance for X */
RUN;


TITLE4 'Script 5d: Adding a common covariance to a full A+S+E for X and Y';
DATA outdata;
      SET indata;
      #1 1 0 y_score
      #2 0 1 x_score
      RUN;
DATA yx_data;
      INFILE outdata;
      INPUT dy dx yx_score;
      RUN;

PROC MIXED DATA=yx_data NOCLPRINT COVTEST METHOD=ML; CLASS famid twinid;
      MODEL yx_score = dy dx / NOINT SOLUTION DDFM=SATTERTH CHISQ;
/* Mean and E covariance as Ve2 */
      RANDOM dy dx / SUBJECT=famid(twinid) TYPE=VC;
/* S variances for Y and X */
RANDOM INTERCEPT dy dx / SUBJECT=famid TYPE=VC;
/* S variances */
      RANDOM weightAC weightAU1 weightAU2 / SUBJECT=famid TYPE=TOEP(1); /*
Common A Covariance */
      RANDOM weightAC*dy weightAU1*dy weightAU2*dy / SUBJECT=famid
TYPE=TOEP(1); /* A Variance for Y */
      RANDOM weightAC*dx weightAU1*dx weightAU2*dx / SUBJECT=famid
TYPE=TOEP(1); /* A Variance for X */
RUN;
```

6. SAS MIXED script changes for the Multiple Relatives models

The SAS PROC MIXED setup permits inclusion of variable pedigree sizes without extensive reconfiguring of the data structure. The only alteration needed for the SAS PROC MIXED program code is to add the names of the new weight variables to the RANDOM statement which codes for the genetic effects. To run the example shown in Table III(b), the command used in the twin model:

```
RANDOM weightAC weightAU1 weightAU2 / SUBJECT=famid TYPE=TOEP(1);
```

becomes

```
RANDOM weightAC weightAU1 weightAU2 weightAU3 weightAU4 weightAU5/
  SUBJECT=famid TYPE=TOEP(1);
```

This code handles families with sib-ships of 1 to 5 in size, with any combination of biological and adopted offspring.

A twin sibling design can be estimated by introducing another component (AM) specific to the MZ pair which allows their remaining genetic variance to be correlated and results in the correct expectation of 1.0 for their genetic correlation (see Table III(c)). This would be estimated in PROC MIXED by substituting this line for the genetic components line:

```
RANDOM weightAC weightAM weightAU1 weightAU2 weightAU3 weightAU4 /
  SUBJECT=famid TYPE=TOEP(1);
```

A special twin environmental component *ST*, orthogonal to the family environmental component S, can be estimated by adding the following line to the PROC MIXED input:

```
RANDOM weightST weightSU1 weightSU2 / SUBJECT=famid TYPE=TOEP(1);
```

ACKNOWLEDGMENTS

This work has been supported by grants from the National Institute on Aging (AG07137) to JJM and the National Institute on Alcohol Abuse and Alcoholism (AA11408) and the National Alliance for Research on Schizophrenia and Depression to CAP. This work was conducted in the summer of 2004 while the authors were both Visiting Scholars at the Harris-Manchester College of Oxford University.

REFERENCES

Arbuckle, J. L., and Wothke, W. (1999). *Amos 4.0 User's Guide*. Chicago, IL: Smallwaters Corp.

Bock, R. D., and Bargmann, R. E. (1966). Analysis of covariance structures. *Psychometrika* **31**(4):507–534.

Boomsma, D., and Molenaar, P. C. M. (1987). Constrained maximum likelihood analysis of familial resemblance of twins and their families. *Acta Genet. Med. Gemelloi*. **36**:29–39.

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

Cattell, R. B. (1953). Research designs in psychological genetics, with special reference to the multiple variance method. *Am. J. Human Genet*. **5**:76–93.

Cattell, R. B. (1960). The multiple abstract variance analysis equations and solutions: for nature–nurture research on continuous variables. *Psychol. Rev*. **67**:353–337.

Cattell, R. B. (1963). The interaction of hereditary and environmental influences. *Br. J. Stat. Psychol*. **16**:191–210.

Cattell, R. B. (1982). *The Inheritance of Personality and Ability: Research Methods and Findings*. New York: Academic Press.

Eaves, L. J., and Gale, J. S. (1974). A method for analyzing the genetic basis of covariation. *Behav. Genet*. **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.

Eaves, L., and Erkanli, A. (2003). Markov Monte Carlo approaches to analysis of genetic and environmental components of human development change and G × E interaction. *Behav. Genet*. **33**(3):279–299.

Fisher, R. A. (1918). The correlations between relatives on the supposition of Mendelian inheritance. *Trans. Roy. Soc. Edinburgh* **52**:399–433.

Fisher, R. A. (1925). The resemblance between twins: a statistical examination of Lauterbach's measurements. *Genetics* **10**:569–579.

Goldstein, H. (1995). *Multilevel Statistical Models*, 2nd edn. New York: Oxford Press.

Guo, G., and Wang, J. (2002). The mixed or multilevel model for behavior genetic analysis. *Behav. Genet*. **32**(1):37–49.

Henderson, N. D. (1975). Gene–evniroment interaction in human behavioral development. In K. W. Schaie, V. E. Anderson, G. E. McClearn, and J. Money (eds.), Developmental human behavior genetics: Nature-nurture redefined. Lexington, MA: Lexington Books, pp. 5–24.

Jinks, J., and Fulker, D. W. (1970). A comparison of the biometrical genetics, MAVA, and classical approaches to the analysis of human behavior. *Psychol. Bull*. **73**:311–349.

Jöreskog, K. G., and Sörbom, D. (1999). LISREL 8.30: *LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language*. Hillsdale, NJ: Scientific Software International.

Li, C. C. (1968). Fisher, Wright, and path coefficients. *Biometrics* **24**: 471–483.

Loehlin, J. C. (1965). Some methodological problems in Cattell's multiple abstract variance analysis. *Psychol. Rev.* **72**:156–161.

Loehlin, J. C. (1978). Heredity-environment analyses of Jenck's IQ correlations. *Behav. Genet.* **8**:415–436.

Loehlin, J. C. (1979). Combining data from different groups in human behavior genetics. In J. R. Royce (ed.), *Theoretical Advances in Behavior Genetics*. Leiden, The Nethderlands: Sijthoff and Noordhoff, pp. 303–334.

Maes, H. M., Neale, M. C., and Eaves, L. J. (1997). Genetic and environmental factors in relative body weight and human adiposity. *Behav. Genet.* **27**:325–351.

Martin, N. G., and Eaves, L. J. (1977). The genetical analysis of covariance structure. *Heredity* **38**:79–95.

Martin, N. G., Eaves, L. J., and Heath, A. (1987). Prospects for detecting genotype × environment interactions in twins with breast cancer. *Acta Genet. Med. Gemelloi.* **36**:5–20.

Mather, K., and Jinks, J. L. (1977). *Introduction to Behavioral Genetics*. Ithaca, NY: Cornell University. Press

McArdle, J. J. (1986). Latent variable growth within behavior genetic models. *Behav. Genet.* **16**(1):163–200.

McArdle, J. J. (1996). A workshop on contemporary behavioral genetic models and methods. *Contemp. Psychol.* **41**(9):886–887.

McArdle, J. J., Connell, J., and Goldsmith, H. H. (1980). Latent variable approaches to measurement, structure, longitudinal stability, and genetic influences: preliminary results from the study of behavioral style. *Behav. Genet.* **10**:609.

McArdle, J. J., Ferrer-Caja, E., Hamagami, F., and Woodcock, R. W. (2002). Comparative longitudinal multilevel structural analyses of the growth and decline of multiple intellectual abilities over the life-span. *Develop. Psychol.* **38**(1):115–142.

McArdle, J. J., and Goldsmith, H. H. (1990). Some alternative structural equation models for multivariate biometric analyses. *Behav. Genet.* **20**(5):569–608.

McArdle, J. J., and Hamagami, F. (1996). Multilevel models from a multiple group structural equation perspective. In G. Marcoulides and R. Schumacker (eds.), *Advanced Structural Equation Modeling Techniques*. Hillsdale, N. J.: Erlbaum, pp. 89–124.

McArdle, J. J., and Hamagami, F. (2003). Structural equation models for evaluating dynamic concepts within longitudinal twin analyses. *Behav. Genet.* **33**(3):137–159.

McArdle, J. J., and Prescott, C. A. (1996). Contemporary models for the biometric genetic analysis of intellectual abilities. In D. P. Flanagan, J. L. Genshaft, and P. L. Harrison (eds.), *Contemporary Intellectual Assessment: Theories, Tests and Issues*. New York: Guilford Press, pp. 403–436.

McArdle, J. J., Prescott, C. A., Hamagami, F., and Horn, J. L. (1998) A contemporary method for developmental-genetic analyses of age changes in intellectual abilities. *Develop. Neuropsychol.* **14**(1):69–114.

Muthén, L. K., and Muthén, B. O. (2004). *Mplus VERSION User's Guide*. Los Angeles CA: Muthén and Muthén.

Neale, M. C., Boker, S. M., Xie, G., and Maes, H. H. (2002). *Mx: Statistical Modeling*, 5th edn. Department of Psychiatry, Virginia Commonwealth University, P.O. Box 980126, Richmond, VA 23298.

Neale, M. C., and Cardon, L. (1992). *Methodology for Genetic Studies of Twins and Families*. Dordrecht, The Netherlands: Kluwer Academic.

Pinherio, J. C., and Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer.

Powers, D. A., and Xie, Y. (2000). *Statistical Methods for Categorical Data Analysis*. New York: Academic Press.

Prescott, C. A. (2004). Using the Mplus computer program to estimate models for continuous and categorical data from twins. *Behav. Genet.* **34**(1):17–40.

Purcell, S., and Sham, P. (2002). Variance components models for gene-environment interaction in twin analysis. *Twin Res.* **5**(6):554–571.

SAS Institute (1999). *SAS/STAT Software, Version 8*. Cary NC: SAS Institute.

Skrondal, A., and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*. Chapman and Hall/CRC.

Wright, S. (1918). On the nature of size factors. *Genetics* **5**:367–374.

Wright, S. (1921). Correlation and causation. *J. Agricult. Res.* **20**:557–585.

Edited by Stacey Cherny