Behavior Genetics, Vol. 34, No. 1, January 2004 (© 2004)

Using the Mplus Computer Program to Estimate Models for Continuous and Categorical Data from Twins

Carol A. Prescott[1]

Received 4 Apr. 2002—Final 29 May 2003

Historically, the focus of behavior genetic research was to obtain estimates of the sources of familial resemblance for a single phenotype. Current research strategies have moved beyond heritability estimates to the search for physiological and behavioral mechanisms by which genetic risk is translated into individual differences in behavior and disease liability. Such research questions often require multivariate designs and complex analytic models, including the analysis of continuous and categorical dependent variables within the same model. Recent advances in computer software for categorical data analysis have increased the tools available for researchers in behavior genetics. This paper describes how to use the Mplus software program (Muthén and Muthén, 1998, 2002) for the analysis of data obtained from twins. Example analyses include two- and five-group twin models for univariate and bivariate continuous and categorical variables. Data on alcoholism and age at first drink drawn from the Virginia Adult Twin Study of Psychiatric and Substance Use Disorders are used to illustrate how Mplus can be used to analyze multiple-category variables, recode and transform variables, select subgroups for analysis, handle subjects with incomplete data, include constraints to ensure non-negative loadings, include model covariates, model sex differences, and test alternative hypotheses about mediation of genetic risk by measured variables.

KEY WORDS: Mplus; structural equation models; twins; categorical data; alcoholism.

INTRODUCTION

Before the 1970s, most behavior genetic analyses were conducted by using qualitative or quantitative methods to compare indices of familial resemblance, such as intraclass correlations, calculated separately for each pairing of relative types (e.g., Cattell, 1960). Jinks and Fulker (1970) pioneered the application to human behavior of model-fitting methods previously developed for the analysis of data from plant and animal breeding studies (e.g., Mather and Jinks, 1971). Eaves, Martin, Heath, and colleagues (e.g., Eaves *et al.,* 1978; Martin and Eaves, 1977) adapted the analysis of covariance structure (Joreskog, 1973) to apply model-fitting methods to the analysis of multivariate twin data using FORTRAN programs for genetic applications based on numerical optimization and integration using NAG. About the same time, Fulker (1978) and others (e.g., Behrman *et al.,* 1978) developed specialized programs for fitting models to twin data that also employed maximum likelihood estimation. These model-fitting methods represented a clear advance over the earlier approaches, but required users to be skilled programmers.

The development of the multiple group version of LISREL (Joreskog and Sorbom, 1977) led a number of investigators to appreciate the possibility of using this commercially available program for the analysis of multivariate kinship data. The first behavior genetic applications of LISREL were by McArdle and Goldsmith (1990; McArdle *et al.,* 1980) for twin data and Cantor (1983; Cantor and Nance, 1981) for twin-family data, followed by Boomsma and Molenaar (1986). Soon afterward, the use of LISREL was popularized through the early International Workshops on Methodology of Twin and Family Studies and a special issue of

[1] Virginia Institute for Psychiatric and Behavioral Genetics, Department of Psychiatry, Virginia Commonwealth University, Richmond, Virginia, 23298-0126. Tel: 804-828-5968. Fax: 804-828-1471. e-mail: cprescot@hsc.vcu.edu

0001-8244/04/0100-0017/0

Behavior Genetics (volume 19(1), 1989). In recent years, the use of the LISREL logic has been expanded by the Mx program developed by Neale *et al.,* (e.g., Neale, 1991; Neale *et al.,* 2002), which has the advantages of being free and developed specifically for use with multivariate kinship data.

The biometric analysis of categorical phenotypes has a parallel history. Before the mid-1970s, categorical variables obtained from twins were typically analyzed using concordance ratios (e.g., Smith, 1974). Model-fitting approaches to categorical data were described by Fulker (1973) and Eaves *et al.* (1978), who combined maximum-likelihood analysis with the "threshold" model described by Falconer (1965) and Edwards (1969). Multivariate categorical data could be analyzed with structural modeling software by first computing sample statistics (e.g., polychoric matrices using PRELIS, Joreskog and Sorbom, 1986) followed by a second stage of analyzing these statistics using models for continuous variables. This two-stage approach had several drawbacks, including the inability to test for differences among groups in thresholds. This was addressed in part by early versions of Mx that accepted contingency tables as input for maximum-likelihood estimation and allowed modeling of thresholds.

Advances in computer processing speed and statistical theory have led to a rapid expansion of approaches to the analysis of categorical data. The development of the commercially available LISCOMP program (Muthén, 1988) allowed direct modeling of continuous and categorical data based on raw input. Despite its potential usefulness for a variety of applications in twin and family research, LISCOMP has not been widely used by behavioral genetic researchers. Prescott *et al.* used LISCOMP to estimate bivariate twin models which incorporated both continuous indicators and categorical diagnostic variables (Prescott, 1991; Prescott and Gottesman, 1990; Prescott and Kendler, 1999a). Waller and Muthén (1992) described the use of LISCOMP for analyzing censored variables from a twin sample.

The Mplus software program created by Muthén and Muthén (1998, 2002) shares similar estimation procedures with LISCOMP, but has attempted a simpler input language. Mplus can use a combination of categorical and continuously scored data, employs raw data input, and obtains rapid convergence of models based on multivariate categorical data.

There are no prior published examples of the use of Mplus with twin or extended pedigree data. The goal of this paper is to demonstrate how Mplus can be used for the analysis of data from twins. This paper is intended to be a starting point for behavior geneticists interested in using Mplus. Seven examples are provided to illustrate models for univariate and bivariate continuous and categorical data. Illustrations include analysis of multiple-category variables, recoding and transforming variables, selection of subgroups, handling of subjects with incomplete data, constraints to ensure non-negative loadings, inclusion of covariates, models for sex differences, and models to test hypotheses about genetic mediation.

METHODS

Example Data

Examples 1–2: Simulated data

Scores for 1000 MZ and 1000 DZ twin pairs were simulated using SAS version 8.0 (SAS Institute, 1999). Data were generated for normally distributed scores, with a mean of 100 and standard deviation of 10 and proportions of variance of $a^2 = .40$, $c^2 = .20$, and $e^2 = .40$. Scores were analyzed as continuous and binary. The binary variables were created based on cutting scores of 100, 110, and 120 (i.e., representing category splits of .50:.50, .83:.17 and .975:.025).

Examples 3–7

These are real data from previously published analyses of alcohol-related variables collected as part of the Virginia Adult Twin Study of Psychiatric and Substance Use Disorders (VATSPSUD), which includes two longitudinal studies of participants from the Virginia Twin Registry (now part of the Mid-Atlantic Twin Registry). The first study is of 2439 women from female-female twin pairs, who were interviewed up to four times between 1988 and 1997. The second is a parallel study of male-male and male-female twin pairs (and members of multiple births containing at least one male) that includes 5091 males and 1723 females interviewed once or twice between 1993 and 1998. Details about ascertainment, subject characteristics, and measures are available elsewhere (e.g., Kendler and Prescott, 1999; Prescott *et al.,* 1999).

Five variables are included in the analyses:

Age at interview ("age" in the input scripts): Subject ages ranged from 18 to 56 years (mean = 36.8, SD = 8.9). (Pairs are correlated >.98 for age at interview so only twin 1's age is used).

Zygosity (zyg): Pairs were classified as identical or fraternal based on a computer algorithm of questionnaire responses, estimated to be >95% accurate based on validation by PCR of 12–16 markers (Kendler and Prescott, 1999). Group codes for complete pairs are: 1 = MZF, 2 = DZF, 3 = MZM, 4 = DZM, 5 = DZO.
Lifetime drinking (abst): 1 = lifetime abstinence, 0 = former or current drinker
Age at drinking onset (onset): among drinkers, the first age at which at least one standard drink (1.5 oz ethanol) was consumed. Coded as 98 among lifetime abstainers.
Alcoholism (dx): coded as 0 = unaffected, 1 = alcohol abuse only, 2 = alcohol dependence (with or without alcohol abuse) based on DSM-IV diagnostic criteria (American Psychiatric Association, 1994).

Statistical Analyses

The analyses are organized into seven examples. The first two use simulated data to demonstrate that the correct parameter values could be recovered using the Mplus program. *Example 1* is a two-group twin model for a continuous variable. *Example 2* is a two-group twin model for a binary variable. The remaining examples use data from the VATSPSUD. The original analyses of these data were conducted using LISCOMP and Mx and are available elsewhere (Prescott and Kendler, 1999a,b; Prescott *et al.,* 1999). *Example 3* shows how to fit a scalar sex limitation model to a continuous variable. *Example 4* shows how to fit a non-scalar sex limitation model to a multiple-category variable. *Example 5* illustrates a bivariate model for a continuous predictor and a binary outcome. *Example 6* shows a model to test whether the genetic variance in a categorical outcome is mediated through a continuous variable. *Example 7* shows how to estimate a bivariate model for two binary variables. The input data for Examples 1 and 2 and the input scripts and complete program output for all the examples are available at http://www.vipbg.vcu.edu/~cprescot/twinmplus.

Models were fit directly to raw data using the Mplus program, version 2.12 (Muthén and Muthén, 2002)[1]. Mplus is a DOS-based program that has a Windows interface, providing editing of scripts and viewing of output. Models for continuous data employed a fitting function using maximum likelihood (ML) estimation, the program default for continuous data. Mplus version 2.12 does not feature ML estimation with models for categorical dependent variables (although this is planned for future releases of Mplus), so the analyses presented here used the default method, weighted least squares with mean and variance-adjusted chi-squares (WLSMV). As with other software for categorical variables, the estimation of variance components assumes the categories are created by the placement of thresholds on an underlying continuum. In psychopathology research, this continuum is often referred to as liability, representing an individual's vulnerability to express a disorder. Further details of estimation can be obtained from the Mplus User's Guide (Muthén and Muthén, 2002) and in technical notes obtainable from the Mplus website.

[1] More information about Mplus can be obtained at http://www.statmodel.com.

Model Specification Using Mplus

Table I displays an Mplus input script for a two-group univariate twin model for a continuously scored variable. A short script of 20 lines is sufficient to specify the model. This efficiency occurs in part because parameters and variables are defined for the first group and the only input required for subsequent groups is specification of values that differ from the first group. The input statements are described briefly below and illustrated more fully in the context of the examples.

Title, Data and Variable Lines

The TITLE line includes a title that will be printed at the top of the first page of the output file. The DATA statement specifies the location and format of the file containing the input data. The program defaults are for the data file to be in the same subdirectory as the input script and the input format to be free.

The VARIABLE lines describe the input variables and include: NAMES—specifies the ordering and names of the input variables; USEVARIABLES—specifies the variables to be analyzed (listed first are input variables and next any new variables created within the script); GROUP—defines group membership labels and value codes for multiple group analyses. It is not necessary to sort the input data file or create separate files for each group, making it easy to use the same input file for multiple models.

Three optional statements can be placed in the VARIABLE section. Subsets of observations can be selected for analysis based on values in the dataset using the USEOBS statement. Input variables (but not grouping variables) can be recoded or transformed using

Table I. MPLUS Input Script for Example 1 with Continuous Data

```
TITLE:     2-group Univariate Twin Model for a Continuous Variable
DATA:      FILE = example1.dat;
VARIABLE:  NAMES = famno zyg y1 y2;
           USEVARIABLES = y1 y2;
           GROUP = zyg(1=mz 2=dz);
ANALYSIS:  TYPE = MEANSTRUCTURE;
MODEL:                                  ! set up values for all groups
  [y1*100] (1); [y2*100] (1);           ! means
  y1@0; y2@0;                           ! fix residual variances to zero
  A1 BY y1*5 (11); A2 BY y2*5 (11);     ! additive genetic loadings
  C1 BY y1*5 (12); C2 BY y2*5 (12);     ! common envt loadings
  E1 BY y1*8 (13); E2 BY y2*8 (13);     ! specific envt loadings
  [A1@0 A2@0 C1@0 C2@0 E1@0 E2@0];      ! fix latent variable means=0
   A1@1 A2@1 C1@1 C2@1 E1@1 E2@1;       ! fix latent variable vars=1
   A1 WITH A2@1; A1 WITH C1-E2@0;    A2 WITH C1-E2@0;    ! latent variable corrs
   C1 WITH C2@1; C1 WITH E1-E2@0;    C2 WITH E1-E2@0;
   E1 WITH E2@0;
MODEL dz:              ! parameters which differ from 1st group
  A1 WITH A2@0.5;
OUTPUT: SAMP RES STAND;
```

the DEFINE statement. Cases with incomplete data can be excluded based on missing values specified in the MISSING statement. These options are illustrated in the examples.

Analysis, Model, and Output Lines

In the ANALYSIS section, MEANSTRUCTURE indicates that the model includes means (or thresholds) as well as covariance structure. The PARAM statement is used with categorical dependent variables to indicate which parameterization is being used (see Appendix 1). H1 indicates that the model is an unrestricted model for the means, variances, and covariances of the observed variables. The likelihood of the H1 model is compared to that of the specified model (H0) to obtain the chi-square test of model fit.

The first MODEL statement specifies the general model that applies to all groups unless specified otherwise. Subsequent model statements must include a group label as listed in the GROUP statement (e.g., DZ) and only need to include specifications for parameters that differ from the general model. In the examples, the general model is written for the first group listed in the GROUP statement (MZ, MZF, or MZM).

If an estimation method other than the default (ML for continuous, WLSMV for categorical variables) is desired, it can be selected using the ESTIMATION statement. Other options include weighted least squares (WLS) and unweighted least squares (ULS).

The OUTPUT command is optional. Default output includes the parameter estimates, their standard errors, and the ratio of estimates divided by standard errors (t values). The output requested in the script in Table I includes sample statistics (SAMP), residual matrices (RES), and standardized parameter values (STAND). The latter is useful because it prints the explained and residual variances. Many other options are described in the Mplus User's Guide. For example, requesting technical output type 1 (TECH1) results in the parameter specifications and starting values being displayed using matrix notation. Mplus employs the "all-y" model of LISREL, similar to three-matrix RAM notation (McArdle and McDonald, 1984) used in many Mx applications. Note that the parameter labels displayed in the parameter specification section of the output are numbered according to their order of appearance in the script and do not correspond to the user-defined labels. Likelihood-based confidence intervals are available by requesting CINTERVAL on the output line. However, note that these apply to the loading and not the squared loading (variance proportion).

Language Conventions

In brief, the Mplus language uses the following conventions (where **x** represents a specific numeric value): @**x** fixes a parameter to **x**; and *__x__ specifies a free parameter with a starting value of **x**. Integers in parentheses serve as parameter labels to indicate which

parameters are fixed to be equal to one another [e.g., all parameters followed by (11) are equated]. A mean or threshold is indicated in Mplus by enclosing the variable name in [], scaling parameters (described below) are indicated by {}, and variances and covariances are indicated when the variable name is not enclosed. Thresholds are indicated by the variable name followed by a $ sign and an integer, with the integer indicating the ordering of multiple thresholds (e.g., y1$1 for the first threshold of the variable y1, y1$2 for the second). Comments are indicated by a preceding exclamation point (!). All statements (except TITLE and comment lines) must end with a semicolon. Input beyond 80 characters will not be read and will result in a warning printed in the program output.

The Mplus input language uses variable labels rather than matrix positions to specify parameters. This makes understanding a script much easier than with other programs that use matrix locations or parameter labels. The keyword WITH is used to indicate covariances among variables (observed or latent). Thus the statement `A1 with A2@0.5;` in the DZ group specifies that the correlation of A1 and A2 is fixed to 0.5.

Latent variables are defined by their relation to input variables, and this produces a compact notation. In Table I, the statement `A1 BY y1*5(11);` defines the variable A1 as latent (because it was not included in the list of input variables specified by USEVARIABLES), the keyword BY specifies that A1 loads on (is indicated by) the observed variable y1, the loading has a starting value of 5, and it is equated to all other parameters with the label (11).

For convenience, several other conventions have been employed in the input scripts presented here but are not required by Mplus. Keywords are listed using uppercase and user input using lowercase. The numbers 1 and 2 after variable names are used to indicate scores for twins 1 and 2, respectively. (Opposite-sex pairs are ordered so that the male is twin 1, otherwise the assignment to twin 1 or twin 2 can be considered random). Integers used to indicate equality constraints are grouped to identify different types of parameters (e.g., the integers 1 and 2 are used for thresholds or means, 11–13 for loadings). For models containing both sexes, the convention is employed of using the values 1–99 for females and adding a constant (100) to indicate the corresponding parameters for males.

The ordering of the statements within a MODEL section is largely arbitrary. One requirement is that a WITH statement that includes a latent variable must come later in the script than the BY statement that defines that latent variable. In the examples, the following ordering has been used: the first few lines specify parameters and constraints for the input variables (means, thresholds, variances, residuals); the next few lines define the latent variables and their relations to the input variables; the last lines specify the means, variances, and covariances of the latent variables.

EXAMPLES

Two-Group Models for Continuous and Categorical Variables

The goal of the first set of analyses was to ensure that the twin model as specified in Mplus accurately recovers the known parameters used to simulate the data. The numeric estimation in Mplus has been adequately documented elsewhere (e.g., Muthén and Muthén, 1998; Muthén *et al.,* 1999), so one example of continuous and categorical data was deemed adequate.

Example 1: Two-Group Model for Continuous Variables

The input script used for analyzing the simulated continuous variables is displayed in Table I and corresponds to Fig. 1. The input data file includes variables for family number, zygosity, and scores for twin 1 and twin 2 (y1 and y2). The GROUP command specifies that zyg is the grouping variable and defines the values of the group codes.

The first MODEL line sets start values for the score intercepts and equates them for twin 1 and 2. The

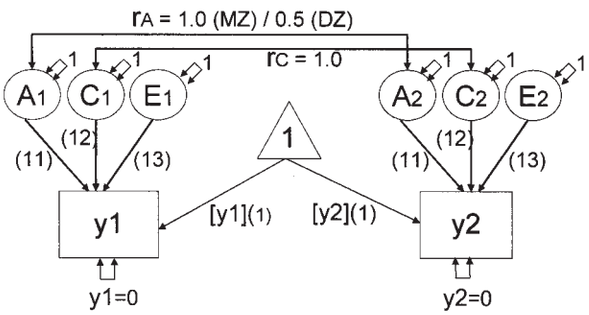

Fig. 1. Univariate twin model for continuous variables as specified with the Mplus program. Numbers in parentheses correspond to the labels used for estimated parameters in the input script for Example 1. The estimated parameters include the observed variable intercepts ([y1] and [y2]), and three sources of variation. Residual variation for the observed scores is fixed to 0 because this is confounded with E variation. Other fixed parameters include the latent variable variances (all fixed to 1) and the correlations between the genetic (r_A) and common environmental (r_C) latent variables.

second line fixes the observed variable residual variances to 0 (e.g., `y1@0;`). The Mplus default is for the residual variances on dependent variables to be free and estimated. They are fixed to 0 here so that the individual-specific variables (E1, E2) are latent variables in the model. For univariate models, e^2 could also be estimated as the residual score variance; the parameterization in Example 1 was selected because it was more familiar to researchers who use twin data.

The next three lines define the latent variables (A, C, E), specify start values for the loadings, and equate the loadings across twin 1 and twin 2. The next two lines fix the latent variable means to 0 and variances to 1. The last three lines specify the values of the correlations among the latent variables. Note that the default is for all of these correlations to be free (not fixed to 0 as in Mx or LISREL). With six latent variables there are 15 correlations to be fixed. These can be written out individually, or can be specified more compactly, as shown in Table I. Note that when a list of variables is specified using a hyphen, the program assumes their order is the same as when they were first defined (in the variable list or with BY statements). For example, C1-E2 signifies C1, C2, E1, and E2, and the statement `A1-A2 WITH C1-E2@0;` fixes eight parameters to 0: A1 and A2 with each of C1, C2, E1, and E2. The other cross-twin correlations are defined for MZ twins as: $r_A = 1.0$ by `A1 WITH A2@1;`, $r_C = 1.0$ by `C1 WITH C2@1;` and $r_E = 0$ by `E1 WITH E2@0;`.

The default for a multiple group model is for most parameters to be invariant across groups. (There are exceptions, such as scaling parameters for categorical variables, described in Example 2). In most cases, the only parameters that need to be specified in subsequent groups are those that differ from the values defined for the general model. In the case of a two-group twin model, only the correlation between the additive genetic latent variables differs between groups. It is defined in the DZ group by the statement `A1 WITH A2@0.5;`.

Selected output for Example 1 is shown in Appendix 1. The format of the output produced by Mplus is very similar to that created by the LISREL and LISCOMP programs. The input statements are repeated (not shown in Appendix 1), and the input data are summarized by listing the number of observations in each group, the number and names of observed and latent variables, and (if requested by including SAMP in the OUTPUT line) the sample statistics for each group, including thresholds or means, and variance-covariance and correlation matrices. Note that all variable names are printed in uppercase in the output file, regardless of how they were formatted in the input script.

The chi-square goodness-of-fit statistic for this model was approximately 4.8, representing twice the difference between the log likelihood value for the null hypothesis of all covariances being 0 (−14524.9), and the value for the alternative hypothesis represented by the specified model (−14522.5). The output includes additional fit indices not shown in Appendix 1, including CFI, TLI, AIC, BIC, and RMSEA.

The MODEL RESULTS section of the output displays model parameters. All fixed parameters are shown at their fixed values (e.g., 0, 1.0, or 0.5) and have standard errors of 0. Estimated parameters have non-0 standard errors. The additive genetic loading (A1 BY Y1) is estimated to be 6.255 (SE = 0.481) and the standardized parameter value is 0.636, shown in the far right column. The common environment loading is estimated at 4.239 (SE = 0.621, stand = 0.431) and the specific environmental loading is 6.298 (SE = 0.138, stand = 0.640). The resulting proportions of variance (obtained by squaring the standardized loadings) are: $a^2 = .404$, $c^2 = .186$, and $e^2 = 0.410$, which sum to 1.0 and are consistent with the observed pair correlations in the simulated data ($r_{MZ} = 0.592$ and $r_{DZ} = 0.387$).

The next few lines of output display the latent variable correlations and means. The Intercepts section shows the mean score across all MZ and DZ subjects is 100.196 (SE = 0.189). Parameter estimates for the DZ group are then displayed, which are identical to those of the MZ group, with the exception of the additive genetic correlation. The complete output next lists any requested output options, including model residuals, model expectations, starting values, the matrix specification of the estimated parameters, and the model run-time.

Submodels (i.e., AE, CE, E only) can be estimated in the usual way. The easiest method to alter the Mplus scripts is to begin with the ACE model and remove parameters by fixing them to 0. For example, to fit the AE model, the line `C1 BY y1*.6(12); C2 BY y2*.6(12);` in Table I would be changed to `C1 BY y1@0;C2 BY y2@0;`. An alternative method is to turn the line that defines the parameter into a comment (using !) so it will be ignored by the program. However, this requires that all subsequent references to the parameter be removed (e.g., all correlations between C1 or C2 and other parameters would now be undefined).

Example 2: Two-Group Model for a Binary Variable

The default Mplus specification for categorical variables is called the DELTA parameterization by the Mplus authors. A binary variable twin model using the DELTA parameterization is portrayed in Fig. 2a, and

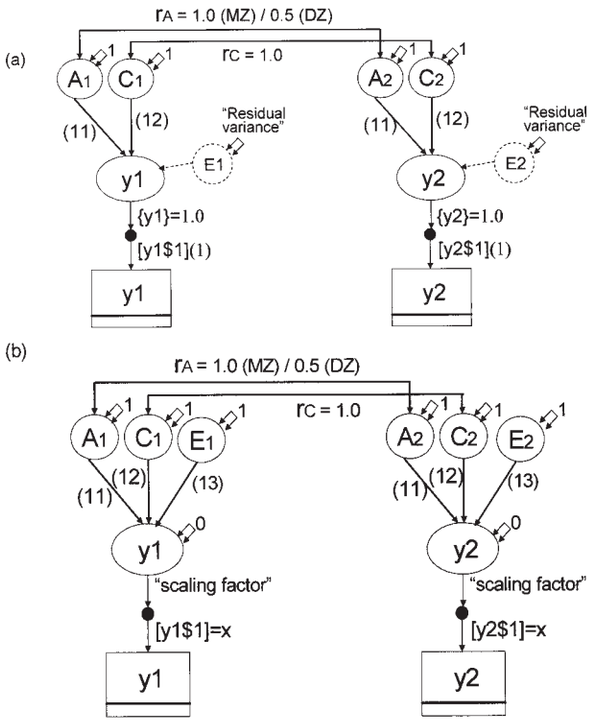

Fig. 2. Mplus specification for a univariate twin model for a categorical variable. The phenotype is a binary variable, indicated by the rectangle with a single horizontal line. The binary variable is transformed into a continuous latent variable using a probit function based on the threshold and scaling factor. (a) The Mplus default, DELTA parameterization, corresponding to the input script in Table II. The thresholds, [y1$1, y2$1], are estimated and the scaling factors, {y1,y2}, are fixed to 1.0. The other estimated parameters are the additive genetic and common environmental loadings. The E factor is shown in dashed lines because (in the DELTA parameterization) it is not an estimated parameter, but is available in the Residual variance section of the program output. (b) Alternate, THETA parameterization which allows the E factors to be modeled. The thresholds, [y1$1, y2$1], are fixed (to the value estimated in a prior run using the DELTA parameterization). The scaling factors are given as part of the program output. Residual variation on the latent scores is fixed to 0 and all three biometric loadings are estimated.

the input script is provided in Table II. The observed binary variable, represented by a box with a single horizontal line, is transformed into a continuous latent variable (of the same name) by a probit function, based on the observed distribution of scores in the categories. In the case of a binary variable, the probit function is based on a constant that is the product of a scaling factor and the threshold. The threshold represents the score on the liability distribution that divides two categories. The model is estimated by standardizing the latent variable (to have mean of 0 and variance of 1) and estimating either the threshold or scaling factor. The dependency of the threshold, scaling factor, and total variance is a mathematical property, not a feature specific to Mplus. It is addressed in LISREL by prior standardization of binary variables (e.g., using PRELIS) and in Mx by including a constraint setting the total variance to unity.

In the script shown in Table II, the threshold is estimated (e.g., `[y1$1](1)`) and the scaling factor is fixed to 1 (e.g., `{y1@1}`). The DELTA parameterization does not allow the residual variance to be estimated as a model parameter. This means standard errors are not available for e^2, but the value of e^2 can be obtained by subtracting all other sources of variation from 1.0 or by referring to the Residual Variance printed in the R-SQUARE section of the program output (see Muthén and Asparouhov [2002] for technical details). An alternative specification (THETA) that includes e^2 as a model parameter is described below and illustrated in Example 7.

For a multiple-category variable, there is additional information from the relative spacing of the thresholds between categories. Most twin model applications will fix the scaling factor to 1 and estimate all the thresholds. However, the scaling factor can be freed to test hypotheses about proportional equality across groups. For example, one might wish to test whether males and females differed in total variance but had equivalent standardized biometric estimates. This could be done with a single parameter test (allowing the scaling parameter to differ from 1.0 in one group) rather than by freeing all the thresholds.

The input for Example 2 differs in five ways from that for Example 1:

1. In the VARIABLE section, the CATEGORICAL statement identifies variables in the analysis that are (or will be transformed to be) categorical. The number of categories in each variable does not need to be specified; the program parses the input data file to determine this.
2. In the Model section, the statement `[y1$1 y2$1](1);` defines the thresholds, assigns the parameter label of 1 and equates the estimates across twin 1 and twin 2. This statement is equivalent to the longer version: `[y1$1] (1); [y2$1](1);` and replaces the one defining the intercepts in the continuous variable script.
3. The statement `{y1@1 y2@1};` fixes to 1 the scaling factors for the observed variables.
4. All statements referring to the latent variables for individual specific variation (E1, E2) have been omitted, including their loadings, means, variances, and correlations.

Table II. MPLUS Input Script for Example 2

```
TITLE:    2-group Univariate Twin Model for a Binary Variable (DELTA param)
DATA:         FILE = example1.dat;
VARIABLE:     NAMES = famno zyg y1 y2;
              USEVARIABLES = y1 y2;
              GROUP = zyg(1=mz 2=dz);
              CATEGORICAL = y1 y2;
DEFINE:       CUT y1-y2 (110);
ANALYSIS:     TYPE = MEANSTRUCTURE;
              PARAM = DELTA;                  ! default method
MODEL:                                        ! set up values for all groups
 [y1$1 y2$1] (1);                             ! thresholds
 {y1@1 y2@1};                                 ! fix scaling factors to 1
 A1 BY y1*.6 (11); A2 BY y2*.6 (11);          ! additive genetic loadings
 C1 BY y1*.3 (12); C2 BY y2*.3 (12);          ! common envt loadings
 [A1-C2@0];                                   ! fix latent variable means=0
  A1-C2@1;                                    ! fix latent variable vars=1
  A1-A2 WITH C1-C2@0;                         ! latent variable correlations
  A1 WITH A2@1; C1 WITH C2@1;
MODEL dz:                                     ! parameters which differ from 1st group
 A1 WITH A2@0.5;
OUTPUT: SAMP RES STAND;
```

5. Example 2 includes an optional DEFINE statement, which categorizes the simulated continuous variables into binary variables based on a cutting score of 110.

Selected output from this model appears in Appendix 2. The sample statistics show the thresholds for the scores of twins 1 and 2 from MZ and DZ pairs. The thresholds correspond to z-scores on a normal distribution and here range from 1.041 to 1.098, which is consistent with selecting the cutting score at +1 SD above the expected mean of the continuous variable. The estimated tetrachoric correlations are $r_{MZ} = .581$ and $r_{DZ} = .323$, consistent with the simulation model values of $r_{MZ} = .600$ and $r_{DZ} = .400$.

The format of the Model results is similar to that for Example 1, except that the output now includes thresholds and scaling parameters for y1 and y2. The estimated threshold pooled across all subjects in the MZ and DZ groups is 1.071 with a standard error of 0.027. Because the score variance is constrained to equal 1, the estimated loadings are already standardized. The additive genetic loading is estimated to be 0.718 (SE = 0.117), which can be squared to obtain the estimated genetic variance (a^2) of 0.516. The common environment loading is estimated to be 0.255 (SE = 0.278), so the c^2 variance is 065. The estimated specific environmental variance can be obtained either by subtraction ($e^2 = 1 - (a^2 + c^2) = 0.419$) or by referring to the residual variance listed in the R-SQUARE section of the output.

Analyses were also conducted for threshold values based on cutting scores of 100 and 120 (i.e., 0 SD and 2 SD). The estimated correlations were identical to those produced by Mx and SAS, and the model parameter estimates were consistent with the correlations.

Use of the THETA Parameterization

Under most circumstances, it is not informative to test whether residual score variance differs from 0, because no measured variables are expected to be error free. However, there may be situations in which obtaining the standard error for residual variance would be desirable. This can be done using an alternative, mathematically equivalent parameterization, called THETA in Mplus (Muthén and Asparouhov, 2002).

In the DELTA parameterization, thresholds are estimated, scaling factors are fixed (to 1), and residual variance is obtained from the estimated model. The THETA parameterization, depicted in Fig. 2b, uses thresholds fixed to an arbitrary value, estimated specific environmental parameter (and associated standard error), calculated scaling factors, and residual variance fixed to 0. The value of the scaling factor is dependent on that chosen for the threshold, and is shown as a calculated parameter in the program output. When the threshold is fixed to any positive value (e.g., *x*), the

standardized parameters are correct as displayed in the output, but the unstandardized parameters need to be rescaled by dividing by the scaling factor. This rescaling can be avoided by fixing the thresholds to their correct values, as obtained by: (1) a prior run using the DELTA parameterization or (2) dividing x by the scaling factor (estimated in a prior run using the THETA parameterization). Fixing the thresholds to their actual values ensures that the standard error and t values (EST/SE), as well as the standardized estimates are correct. An Mplus script corresponding to Fig. 2b is available from the author's website. The THETA input is employed and discussed further in Example 7. Appendix 3 summarizes the differences in model input for continuous variables and categorical variables using the THETA and DELTA parameterizations.

Five-Group Twin Models for Continuous and Categorical Univariate Variables

The next two examples illustrate the use of Mplus to analyze sex differences in continuous and multicategory dependent variables. Data from the Virginia adult twin studies described previously are used to illustrate analyses that account for features of real data, including cases with incomplete data, the inclusion of covariates, recoding of variables, and analysis of a subset of subjects.

Example 3: Model to Test Scalar Sex Differences for a Continuous Variable

Table III shows the Mplus input script for a continuous variable, age at drinking onset,[2] analyzed in all five zygosity groups. The model, portrayed in Fig. 3, can be used to test hypotheses about proportional variances across sex. In this model (sometimes called a scalar sex difference model), males and females have proportional biometric structure. This is specified by males and females having equal biometric loadings but differing in the scalar parameter (e.g., `O1 BY onset1`), which is fixed to 1.0 for one sex (here, females) and estimated for the other sex. Example 3 also illustrates the analysis of a sample with incomplete data and transformation of a dependent variable before analysis.

The inclusion of MISSING and H1 on the ANALYSIS line requests the inclusion of observations with incomplete data, a feature available in Mplus version 2.12 only when all dependent variables are continuous. The missing data estimation uses full information ML (FIML) based on the assumption of missing at random (MAR). This allows data to be used from 526 individuals whose cotwins have known zygosity but are missing onset data, as well as from the 3372 pairs with complete data for both twins. The MISSING statement specifies that missing values for onset are indicated by the values 98 and 99 in the input file.

Onset age is positively skewed, and it might be desirable to transform it before analysis. The DEFINE statement transforms the onset variables by taking the base-10 logarithm and multiplying by 10. (Multiplication by 10 has no effect on the fit, but makes the estimates somewhat easier to work with, because the score variances are $\sim$1.0 rather than $\sim$.01.)

Selected output is displayed in Appendix 3A.[3] To save space, only the parameter estimates for the DZO group are shown. Variables for twin 1 are for males and those for twin 2 are for females. The output first summarizes the proportion of data present. Among DZO pairs, 96.2% of the male twins and 91.6% of the female twins have data, and complete data is available for 87.9% of the pairs. The next section of the displayed output shows the sample statistics, and the final section presents the parameter values.

The square of the scalar parameter in males ($1.089^2 = 1.19$) indicates that the total variance in males is 1.19 times greater than that for females. The variance proportions based on this model are easily calculated from the standardized loadings to be $a^2 = .40$, $c^2 = .08$, and $e^2 = .52$. The sexes differ in their means, estimated as 12.4 (se = .02) for females and 11.9 (se = .02) for males (which correspond to the untransformed values of 17.3 and 15.5 years, respectively). These estimates are listed in the program output under the INTERCEPTS heading because they are the expected scores after partialing the effects of other variables in the model. In this example, the intercepts and score means are identical because all the variables contributing to onset (A, C, E) have means of 0. When the intercepts and score means differ, the expected score means can be obtained from the Estimated Model and Residuals portion of the output, requested with the RES command.

[2] Age at drinking onset would be more properly modeled as a duration variable, to account for the right censoring among nondrinkers who may eventually use alcohol. In the original publication based on the drinking onset data (Prescott and Kendler, 1999a), we included analyses to test for the influence of such censoring, but found little effect, probably because >95% of the sample had used alcohol. Event history models for twin data have been presented by Meyer *et al.* (1991) and Yashin *et al.* (1999), among others.

[3] Appendices 3A, 4–7, all input scripts, and output are available at http://www.vipbg.vcu.edu/~cprescot/twinmplus.

Table III. MPLUS Input Script for Example 3

```
TITLE:   5-Group Model for Transformed Age at Drinking Onset, MF proportional
DATA:    FILE = onset99.dat;
VARIABLE: NAMES = famno mpair zyg age dx1 abst onset1 dx2 abst2 onset2;
     USEVARIABLEs = onset1 onset2;
     GROUP = zyg(1=mzf 2=dzf 3=mzm 4=dzm 5=dzo); ! excludes prs with unk zyg
     MISSING = onset1-onset2(98,99);
DEFINE:  onset1=10*(log10(onset1));  onset2=10*(log10(onset2));
ANALYSIS: TYPE = MEANSTRUCTURE MISSING H1;
MODEL:
  [onset1 onset2] (1);                  ! means
  onset1@0; onset2@0;                   ! fix resid variance=0
  O1 BY onset1@1; O2 BY onset2@1;       ! scalars
! Biometric loadings
 A1 BY O1*.2 (11); A2 BY O2*.2 (11);    ! additive genetic loadings
 C1 BY O1*.2 (12); C2 BY O2*.2 (12);    ! common envt loadings
 E1 BY O1*.8 (13); E2 BY O2*.8 (13);    ! indvl envt loadings
! Means, variances & correlations for latent variables
  [O1-E2@0]; !fix latent variable means
   O1-O2@0; A1-E2@1;                    !fix latent variable variances
   O1-O2 WITH A1-E2@0;                  ! fix correlations
   A1-A2 WITH C1-E2@0;
   C1-C2 WITH E1-E2@0;
   O1 WITH O2@0; A1 WITH A2@1; C1 WITH C2@1; E1 WITH E2@0;
MODEL dzf:
  A1 WITH A2@0.5;
MODEL mzm:                              ! in groups mzm, dzm and dzo, need to specify params
  [onset1 onset2] (101);                ! which differ from values set up in the first group
  O1 BY onset1*1.2 (102); O2 BY onset2*1.2 (102);
MODEL dzm:
  [onset1 onset2] (101);
  O1 BY onset1*1.2 (102); O2 BY onset2*1.2 (102);
  A1 WITH A2@0.5;
MODEL dzo:
  [onset1] (101);
  O1 BY onset1*1.2 (102);
  A1 WITH A2@0.5;
OUTPUT: SAMP STAND RES TECH1;
```

As with other programs, various hypotheses about sources of sex differences can be tested by comparing the fits of alternative models. To test equality (versus proportionality) of biometric estimates across sex, the scalar could be fixed to 1.0 in males by specifying `O1 BY onset1@1; O2 BY onset2@1;` in the MZM, DZM, and DZO groups. (Note that because the Mplus default is to hold factor loadings invariant across groups, the same result would be obtained by commenting out or deleting the lines freeing this parameter in the male groups.) The fit of a model testing male-female equality of variance components was $\chi^2 = 96.5$, $df = 20$, much worse than the fit of the scalar model used in Example 3 ($\chi^2 = 69.4$, $df = 19$).

A model to test invariance of means and loadings could be specified by taking the previous equality model and changing the parameter label for the mean among males (101) to be the same as that among females (1). This model obtained a fit of $\chi^2 = 489.9$, $df = 21$, reflecting the substantial sex differences in age at drinking onset.

The improved fit of a nonscalar sex differences model, allowing the sexes to differ in all parameters (means, total variance, and sources of variance) was $\Delta\chi^2 = 51.9, df = 17$, suggesting the variance structure

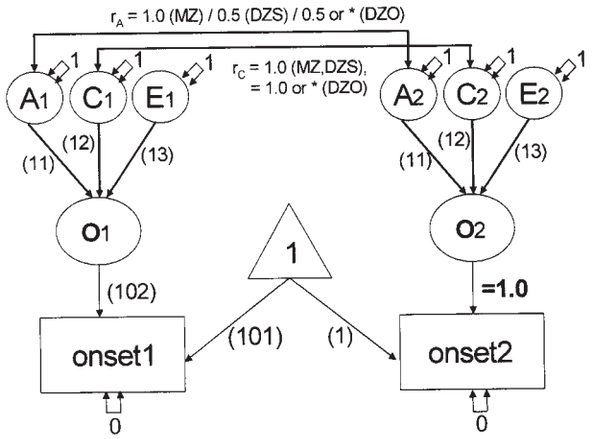

Fig. 3. Mplus specification for estimating a scalar sex differences model for a continuous variable (Example 3). The parameters as shown apply to opposite-sex pairs. Proportional equality of variance across sexes is specified by allowing males and females to have equal biometric loadings and multiplying the total variance in one sex by a scalar. Males and females are allowed to have different score intercepts. The genetic covariance for opposite-sex pairs may be fixed to 0.5 or free (*); the common environmental covariance may be fixed to 1.0 or free (*).

differs for the two sexes. Mplus input for a non-scalar sex differences model is provided in Example 4.

Example 4: Five-Group Sex-Difference Model for a Categorical Variable with Covariates

Table IV and Fig. 4 show input used to test for nonscalar sex differences in the sources of variance for a three-category variable. Mplus employs casewise deletion with categorical outcomes, so this analysis includes only twin pairs with complete diagnostic information. The model was fit to DSM-IV alcoholism diagnoses categorized into unaffected = 0, alcohol abuse only = 1, and alcohol dependence = 2. A three-category variable requires two thresholds. These are indicated in the input script by $1 and $2, (i.e., the labels dx1$1 and dx1$2 represent the first and second thresholds for the variable dx1). Males and females are allowed to have different thresholds for diagnosis (parameter labels 1 and 2 for females, 101 and 102 for males) and different A and C loadings (11 and 12 for females, 111 and 112 for males). The male-female genetic correlation is estimated by specifying `Adx1 WITH Adx2*0.5;` in the DZO group.

In this model, the biometric components of variance are estimated after partialing the linear and quadratic effects of age. This is done because in this (and other) samples, there are age cohort effects on the prevalence of alcoholism. A new variable, dage, is created using the DEFINE statement by centering age around the approximate mean in the data (35 years) and rescaling by 10. This is done to aid interpretation of the regression coefficients (centering places the 0 point within the data and rescaling puts the variance in a similar metric as that for diagnosis). A second DEFINE statement creates a quadratic age variable (dagesq). The default in Mplus is for the covariances between exogenous variables (those influencing other observed variables) and latent variables to be free. Here, these covariances are all fixed to 0 by the statement `dage-dagesq WITH Adx1-Cdx2@0;`. Regressions among observed variables are specified using the ON keyword (e.g., `dx1 ON dage*-.3(21);`). In addition to the regression parameters (21 and 22 for females, 121 and 122 for males), the script includes parameters for the means of dage and dagesq (23 and 24), their variances (25, 26), and their covariance (27). The means, variances, and covariances of the age variables are equated over groups, because all the zygosity groups are assumed to be drawn from the same population with respect to age. The Mplus default is for the means of exogenous variables to be held equal over groups, but not the variances and covariances, so these constraints are specified in the MODEL statements for each group.[4]

Appendix 4 displays the parameter estimates for the DZO group obtained from fitting this model. The thresholds for diagnosis were estimated for males as: 0.223 for the cutpoint between unaffected and alcohol abuse and 0.591 for the cutpoint between abuse and dependence, corresponding to prevalences of 13.4% with abuse only (i.e., the proportion of a normal distribution between z-scores of .223 and .591) and 27.8% meeting criteria for lifetime alcohol dependence (i.e., the proportion of a normal distribution above z = .591). The thresholds for females were 0.929 and 1.213, corresponding to prevalences of 6.4% with abuse only and 11.1% with alcohol dependence. The estimated tetrachoric correlations for the three-category diagnosis in the pairs were: MZF = .533, DZF = .170, MZM = .525, DZM = .281, DZO = .116.

[4] Note that modeling the age parameters (by equating them across twins or groups) has the effect of bringing them into the model—assuming joint normality for the outcome variable as well as the covariate—and invokes the biserial/tetrachoric correlation approach to estimation. When covariates are not modeled, Mplus uses a probit regression-based estimation approach that only assumes normality of the distribution of the outcomes given the covariate and makes no assumptions about the distribution of the covariate. Although the latter approach may be preferable statistically, it may not be practical for twin models where one wants to constrain equality of the regressions across twins or groups.

Table IV. MPLUS Input Script for Example 4

```
TITLE:  5-group Model for 3-Category Diagnosis with Age Regressions & Free Rg
DATA:       FILE = onset99.dat;
VARIABLE:   NAMES = famno mpair zyg age dx1 abst1 onset1 dx2 abst2 onset2;
            USEVARIABLEs = dx1 dx2 dage dagesq;
            CATEGORICAL = dx1 dx2;
            GROUP = zyg(1=mzf 2=dzf 3=mzm 4=dzm 5=dzo);
            MISSING = ALL(98,99);
DEFINE:     dage = (age-35)/10; dagesq = (dage*dage);
ANALYSIS:   TYPE = MEANSTRUCTURE;
MODEL:
  {dx1@1 dx2@1};                                ! scaling
  [dx1$1 dx2$1] (1); [dx1$2 dx2$2] (2);         ! thresholds
  dx1 ON dage*-.3 (21);   dx2 ON dage*-.3 (21); ! age regressions
  dx1 ON dagesq*-.1 (22); dx2 ON dagesq*-.1 (22);
  [dage](23); [dagesq] (24);                    ! estimate means
  dage(25); dagesq (26);                        ! variances &
  dage WITH dagesq (27);                        ! covariance of covariates
! Biometric loadings
  Adx1 BY dx1*.6 (11); Adx2 BY dx2*.6 (11);  ! additive genetic loadings
  Cdx1 BY dx1*.6 (12); Cdx2 BY dx2*.6 (12);  ! common envt loadings
! Means, variances & correlations for latent variables
  [Adx1-Cdx2@0];                                ! fix means=0
  Adx1-Cdx2@1;                                  ! fix vars=1
  dage-dagesq WITH Adx1-Cdx2@0;
  Adx1-Adx2 WITH Cdx1-Cdx2@0;
  Adx1 WITH Adx2@1; Cdx1 WITH Cdx2@1;
MODEL DZF:                                      ! PARAMETERS WHICH DIFFER FROM FIRST GROUP (MZF)
  dage(25); dagesq (26);                        ! equate variances &
  dage WITH dagesq (27);                        ! covariance of covariates
  Adx1 WITH Adx2@0.5;
MODEL MZM:
  [dx1$1 dx2$1] (101);   [dx1$2 dx2$2] (102); ! thresholds for males
  dage(25); dagesq (26);                        ! variances &
  dage WITH dagesq (27);                        ! covariance of covariates
  dx1 ON dage*.3 (121);   dx2 ON dage*.3 (121);
  dx1 ON dagesq*.1 (122); dx2 ON dagesq*.1 (122);
  Adx1 BY dx1*.6 (111); Adx2 BY dx2*.6 (111); ! loadings for males
  Cdx1 BY dx1*.6 (112); Cdx2 BY dx2*.6 (112);
MODEL DZM:
  [dx1$1 dx2$1] (101);  [dx1$2 dx2$2] (102);
  dx1 ON dage*.3 (121);   dx2 ON dage*.3 (121);
  dx1 ON dagesq*.1 (122); dx2 ON dagesq*.1 (122);
  dage(25); dagesq (26);                        !   variances &
  dage WITH dagesq (27);                        !   covariance of covariates
  Adx1 BY dx1*.6 (111); Adx2 BY dx2*.6 (111);
  Cdx1 BY dx1*.6 (112); Cdx2 BY dx2*.6 (112);
  Adx1 WITH Adx2@0.5;
MODEL DZO:                                      !Only need to specify male parameters
  [dx1$1] (101); [dx1$2] (102);
  dx1 ON dage*.3 (121);
  dx1 ON dagesq*.1 (122);
  dage(25); dagesq (26);                        !   variances &
  dage WITH dagesq (27);                        !   covariance of covariates
  Adx1 BY dx1*.6 (111);
  Cdx1 BY dx1*.6 (112);
  Adx1 WITH Adx2*0.5;
OUTPUT: SAMP STAND RES TECH1;
```

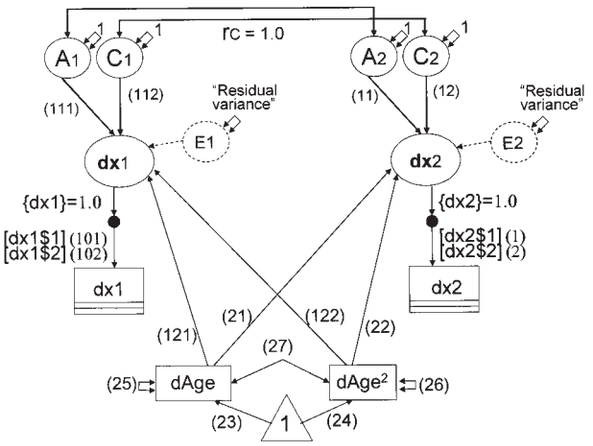

Fig. 4. Mplus specification for a nonscalar sex differences model with a three-category outcome variable (shown as a box with two horizontal lines), including linear and quadratic age as covariates (Example 4). The parameters as shown apply to opposite-sex pairs. Estimated parameters are indicated by integers in parentheses.

The age regressions for males and females indicate significant negative quadratic effects (male = −.133 and female = −.088), such that the oldest and youngest cohorts have lower prevalences of alcoholism. For women, there is also a significant negative linear effect (−.149), indicating the prevalences are lowest in the oldest age-group. Although statistically significant in this large sample, the age effects account for a very small proportion of the overall variance in diagnosis liability.

The variance components for males are: $a^2 = .491$, $c^2 = .019$, $e^2 = .475$, dage = .000, and dagesq = .013, and for females: $a^2 = .483$, $c^2 = .000$, $e^2 = .488$, dage = .017, and $dage^2 = .006$. The total variance in this model includes age effects, so simply squaring the values in the StdYX column for a, c, and e would not produce the correct standardized parameter estimates. Instead, the estimates would need to be standardized as a proportion of the total variance after partialing age (i.e., 0.985 for males and 0.971 for females).

The estimated male-female genetic correlation of .204 is not significantly different from the default value of .50. When the model was run with the genetic correlation fixed to .50, the common environment loading in females became slightly negative. In some circumstances with five-group models, a negative loading in one sex and a positive loading in the other could indicate over-fitting to the data. Mplus does not permit the user to specify parameter boundaries. However, a specification in which the square roots of the genetic and common environmental loadings are estimated will constrain the male-female covariance to be non-negative (see Appendix 8 and Fig. 8).

Bivariate Models for Categorical Data

Examples 5, 6, and 7 illustrate several approaches to estimating the sources of covariation between drinking onset age and liability for alcoholism. To save space, Tables V–VII display program input for two-group models, but could be expanded to five groups as with Examples 3 and 4.

Example 5: Estimating the Sources of Covariation Between a Continuous and a Categorical Variable

The first model, depicted in Fig. 5 for one individual, estimates the sources of covariation between drinking onset (treated as a continuous score) and liability for alcoholism based on a Cholesky decomposition. Liability to alcoholism is regressed on the biometric factors contributing to onset, and there are also sources of residual variation for diagnosis. The sources of covariation between diagnosis and onset are represented by the loadings labeled 51, 52, and 53. The model is fit to data from female twin pairs in which both twins are nonabstainers.

The input script corresponding to Fig. 5 is presented in Table V. The input illustrates some useful Mplus features for selecting cases and recoding variables. Observations are limited to female MZ and DZ pairs in which both twins are nonabstainers by the statement `USEOBS = (abst1==0 and abst2==0) and (zyg==1 or zyg==2);`. (The double equal symbol corresponds to the logical operation "is equal to".) The DEFINE command is used to recode the three-category diagnosis variable into two categories. Values of 1 or 2 (i.e., greater than the 0 cutting score) are recoded so a score of 1 now corresponds to having either alcohol abuse or alcohol dependence. The next DEFINE statement recodes diagnosis so affected individuals have value 0 and unaffected have value 1. This is done for convenience so that the covariance between diagnosis and younger drinking age will be estimated as positive rather than negative.

Selected output is shown in Appendix 5. The cross-twin cross-variable correlations (onset1-dx2, onset2-dx1) are more similar for MZ than DZ pairs, and among MZ pairs are about as large as the within-person cross-variable correlations (dx1-onset1 and dx2-onset2), suggesting genetic covariance. The proportions of variance (obtained by squaring the standardized loadings) in onset are $a_o^2 = (.639)^2 = .408$, $c_o^2 = (.355)^2 = .126$, and $e_o^2 = (.683)^2 = .466$. The sources of onset-dx covariation are estimated to be $a_{do}^2 = (0.667)^2 = .445$, $c_{do}^2 = (-0.064)^2 = .004$, and $e_{do}^2 = (0.047)^2 = .002$. Thus 45% of the variance in alcoholism liability is estimated as shared with that in

Table V. MPLUS Input Script for Example 5

```
TITLE:  2-Group Bivariate Model for Age at Drinking Onset and Diagnosis
DATA:     FILE = onset99.dat;
VARIABLE:  NAMES = famno mpair zyg age dx1 abst1 onset1 dx2 abst2 onset2;
       USEVARIABLES = dx1 dx2 onset1 onset2;
       CATEGORICAL = dx1 dx2;
       USEOBS = (abst1==0 and abst2==0) and (zyg==1 or zyg==2);
       GROUP = zyg(1=mzf 2=dzf);
       MISSING = dx1,abst1,dx2,abst2(99) onset1,onset2(98,99);
DEFINE: CUT dx1(0);  CUT dx2(0);                    ! for DSM4 abuse/dependence
DEFINE: dx1=1-dx1;   dx2=1-dx2;                     ! recoding affected=0, unaff=1
ANALYSIS: TYPE = MEANSTRUCTURE;
MODEL:
  {dx1@1 dx2@1};                                    ! scaling
  [dx1$1 dx2$1] (1);                                ! thresholds
  [onset1 onset2] (11);                             ! means
  onset1@0 onset2@0;                                ! residual variances on onset
!BIOMETRIC COMPONENTS FOR ONSET
   Ao1 BY onset1*.6 (41); Ao2 BY onset2*.6 (41);
   Co1 BY onset1*.6 (42); Co2 BY onset2*.6 (42);
   Eo1 BY onset1*.8 (43); Eo2 BY onset2*.8 (43);
  [Ao1-Eo2@0];
   Ao1-Eo2@1;
   Ao1 BY dx1*.6 (51); Ao2 BY dx2*.6 (51);
   Co1 BY dx1*.6 (52); Co2 BY dx2*.6 (52);
   Eo1 BY dx1*.6 (53); Eo2 BY dx2*.6 (53);
!RESIDUAL COMPONENTS FOR DX
   Adx1 BY dx1*.6 (61); Adx2 BY dx2*.6 (61);
   Cdx1 BY dx1*.6 (62); Cdx2 BY dx2*.6 (62);
  [Adx1-Cdx2@0];
   Adx1-Cdx2@1;
!CORRELATIONS AMONG BIOMETRIC COMPONENTS
  Ao1-Ao2 WITH Co1-Cdx2@0;
  Co1-Co2 WITH Eo1-Cdx2@0;
  Eo1-Eo2 WITH Adx1-Cdx2@0;
  adx1-adx2 WITH Cdx1-Cdx2@0;
  Ao1 WITH Ao2@1; Adx1 WITH Adx2@1;
  Co1 WITH Co2@1; Cdx1 WITH Cdx2@1;
  Eo1 WITH Eo2@0;
MODEL dzf:
   Ao1 WITH Ao2@0.5;  Adx1 WITH Adx2@0.5;
OUTPUT: SAMP STAND RES TECH1;
```

onset, and this is almost entirely due to additive genetic covariance. (Although the c loading is negative, it is essentially 0. If it were larger, the model could be fit with this loading fixed to 0 or using a square-root parameterization.) The estimates for residual diagnosis liability are virtually 0 ($a_d = -.005$, $c_d = .000$). This means that all the familial variation in diagnosis is estimated as covarying with the additive genetic sources for onset. The remaining 55% is nonfamilial residual variation, reflected in the Residual variance column of the R-SQUARE portion of the output.

Example 6: Mediation Model for Sources of Covariation

In this model the covariation is modeled through the onset variable. This contrasts with the less restrictive model used for Example 5 that tests whether onset and age have common genetic and/or environmental sources. The fit of this mediation model compared to the prior model represents a test of whether the phenotypic variable can be considered a direct risk factor for the outcome (e.g., Prescott *et al.,* in press). To create the script for the mediation model, the regressions of

Table VI. MPLUS Input Script for Example 6

```
TITLE:  2-group Bivariate Mediation Model for diagnosis and drinking onset
          with fixed unreliability for onset among female drinking pairs
DATA:       FILE = onset99.dat;
VARIABLE:   NAMES = famno mpair zyg age dx1 abst1 onset1 dx2 abst2 onset2;
            USEVARIABLEs = dx1 dx2 onset1 onset2;
            CATEGORICAL = dx1 dx2 ;
            USEOBS = (abst1==0 and abst2==0) and (zyg==1 or zyg==2);
            GROUP = zyg(1=mzf 2=dzf);
            MISSING = dx1,abst1,dx2,abst2(99) onset1,onset2(98,99);
DEFINE:     CUT dx1(0);  CUT dx2(0); dx1=1-dx1; dx2=1-dx2;
ANALYSIS:   TYPE = MEANSTRUCTURE;
MODEL:
 {dx1@1 dx2@1};                                  !scaling
 [dx1$1 dx2$1]  (1);                             ! thresholds
 [onset1 onset2]  (11);                          !means
 onset1@4.502 onset2@4.502;                      !estimated unreliability = .395*total variance
!BIOMETRIC COMPONENTS FOR ONSET
 Ao1 by onset1*.6 (41); Ao2 by onset2*.6 (41);
 Co1 by onset1*.6 (42); Co2 by onset2*.6 (42);
 Eo1 by onset1*.8 (43); Eo2 by onset2*.8 (43) ;
[Ao1-Eo2@0];
 Ao1-Eo2@1;
! MEDIATION PARAMETERS
 dx1 on onset1*.6 (54); dx2 on onset2*.6 (54);
! RESIDUAL COMPONENTS FOR dx
 Adx1 by dx1*.6 (61); Adx2 by dx2*.6 (61);
 Cdx1 by dx1*.6 (62); Cdx2 by dx2*.6 (62);
[Adx1-Cdx2@0];
 Adx1-Cdx2@1;
!CORRELATIONS AMONG BIOMETRIC COMPONENTS
  Ao1-Ao2 with Co1-Cdx2@0;
  Co1-Co2 with Eo1-Cdx2@0;
  Eo1-Eo2 with Adx1-Cdx2@0;
  Adx1-Adx2 with Cdx1-Cdx2@0;
  Ao1 with Ao2@1; Adx1 with Adx2@1 ;
  Co1 with Co2@1; Cdx1 with Cdx2@1 ;
  Eo1 with Eo2@0;
 MODEL dzf:
  Ao1 with Ao2@0.5; Adx1 with Adx2@0.5;
 OUTPUT: SAMP STAND RES TECH1;
```

diagnosis on the latent variables underlying onset (parameters 51–53 in Example 5) could simply be replaced by a regression of diagnosis on onset. A conceptual limitation of this approach is that the common E factor (e.g., E_{OI}) in a bivariate model contains the unreliable variation in the predictor variable, whereas E covariation is assumed to be estimated without error. Thus a better model, shown in Fig. 6, estimates the regression of diagnosis on the predictor after partialing the unreliable variance in the predictor. Although the unreliable variance cannot be estimated based on a single measure of the predictor, an external estimate can be employed (e.g., from published norms, a test-retest estimate, or another sample). To allow a fair test of the model, the unreliability estimate should be chosen such that the resulting reliable variance does not exceed the variation in the predictor that overlaps with the outcome (as estimated using the Cholesky version of the bivariate model). In Example 6, the reliability estimate is obtained from the short-term test-retest correlation of reported drinking onset obtained in this sample, $r = .605$. The unreliable variance (u^2 in Fig. 6) is the product of the unreliability (i.e., $1 - .605 = .395$) and the total score variance. The sample variance can be

Table VII. MPLUS Input Script for Example 7

```
TITLE:      2-Group Bivariate Model for Diagnosis and Categorized Onset Age
DATA:       FILE = onset99.dat;
VARIABLE:   NAMES = famno mpair zyg age dx1 abst1 onset1 dx2 abst2 onset2;
            USEVARIABLEs = dx1 dx2 onset1 onset2;
            CATEGORICAL  = dx1 dx2 onset1 onset2 ;
            GROUP = zyg(3=mzm 4=dzm);
            MISSING = ALL(99);
DEFINE:     CUT dx1(0); CUT dx2(0);                          ! for dsm4 abuse/dependence
DEFINE:     CUT onset1(15); CUT onset2(15);                  ! recodes to binary variable
DEFINE:     onset1=1-onset1;  onset2=1-onset2;               ! recoding to early=1 later=0
ANALYSIS:   TYPE = MEANSTRUCTURE; PARAM = THETA;
MODEL:
 [dx1$1@1 dx2$1@1 onset1$1@1 onset2$1@1];                    ! values estimated in prior run
!BIOMETRIC COMPONENTS FOR ONSET
 Ao1 BY onset1*.6 (41); Ao2 BY onset2*.6 (41);
 Co1 BY onset1*.6 (42); Co2 BY onset2*.6 (42);
 Eo1 BY onset1*.8 (43); Eo2 BY onset2*.8 (43);
[Ao1-Eo2@0];
 Ao1-Eo2@1;
!BIOMETRIC COMPONENTS FOR dx
 Ao1 BY dx1*.6 (51); Ao2 BY dx2*.6 (51);
 Co1 BY dx1*.6 (52); Co2 BY dx2*.6 (52);
 Eo1 BY dx1*.6 (53); Eo2 BY dx2*.6 (53);
 Adx1 BY dx1*.6 (61); Adx2 BY dx2*.6 (61);
 Cdx1 BY dx1*.6 (62); Cdx2 BY dx2*.6 (62);
 Edx1 BY dx1*.6 (63); Edx2 BY dx2*.6 (63);
[Adx1-Edx2@0];
 Adx1-Edx2@1;
!CORRELATIONS AMONG BIOMETRIC COMPONENTS
 Ao1-Ao2 WITH Co1-Edx2@0;     Ao1 WITH Ao2@1;
 Co1-Co2 WITH Eo1-Edx2@0;     Co1 WITH Co2@1;
 Eo1-Eo2 WITH Adx1-Edx2@0;    Eo1 WITH Eo2@0;
 Adx1-Adx2 WITH Cdx1-Edx2@0;  Adx1 WITH Adx2@1;
 Cdx1-Cdx2 WITH Edx1-Edx2@0;  Cdx1 WITH Cdx2@1;  Edx1 WITH Edx2@0;
MODEL MZM:
 dx1@0 dx2@0 onset1@0 onset2@0;                               ! fixing residual variance=0
MODEL DZM:
 dx1@0 dx2@0 onset1@0 onset2@0;
 Ao1 WITH Ao2@0.5; Adx1 WITH Adx2@0.5;
OUTPUT: SAMP STAND RES TECH1;
```

calculated within the model, but this requires a complex script with extra latent variables, so it is simpler to enter the unreliability as a fixed parameter.

Input for the model depicted in Fig. 6 is shown in Table VI and selected output in Appendix 6. The loadings for onset are similar to those from Example 5, except that to obtain the total individual-specific variance in onset, one would square the E loading (0.732) and add the fixed estimate of unreliable variance, shown in the Residual Variances line (4.502). The standardized proportions of variance for onset are $a_o^2 = (.666)^2 = .444$, $c_o^2 = (.338)^2 = .114$, $e_o^2 = (.217)^2 = .047$, and $u_o^2 = .395$. Because the unreliability for onset is fixed, the R-SQUARE section of the output shows the explained variance for onset as the remainder, .605. The proportion of variance in diagnosis that overlaps with onset is obtained by squaring the standardized regression coefficient (from the StdYX column, .464). This yields .215, a value substantially lower than the estimated overlap of .451 obtained in Example 5. The remaining (non–onset related) sources of variance in diagnosis include $a_d^2 = (.568)^2 = .323$, $c_d^2 = (.000)^2 = .000$, and residual variance of .461.

The WLSMV goodness-of-fit statistic for this model is 34.5 for 10 parameters, compared with a fit of 22.6 for 8 parameters for the standard covariance

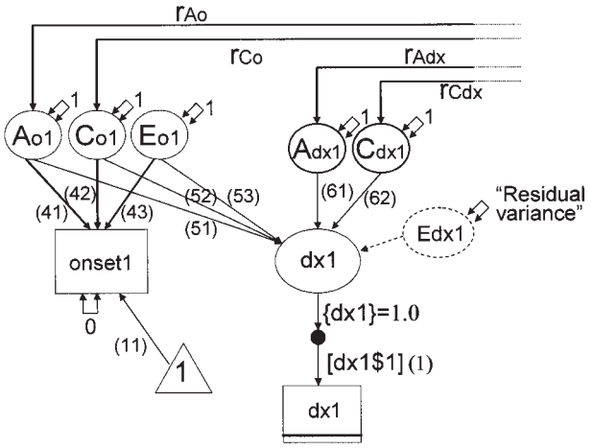

Fig. 5. Mplus specification for a bivariate Cholesky decomposition model for estimating covariation between a continuous predictor variable and a binary outcome (Example 5). The model is shown for one individual (twin 1). One set of factors (Ao, Co, Eo) contributes to variation in both onset age and liability to alcoholism diagnosis (dx). A set of residual factors contributes to the remaining variation in alcoholism liability. Individual-specific diagnosis residual variance is not estimated as an independent parameter, but is available as a calculated value in the Residual variance section of the program output.

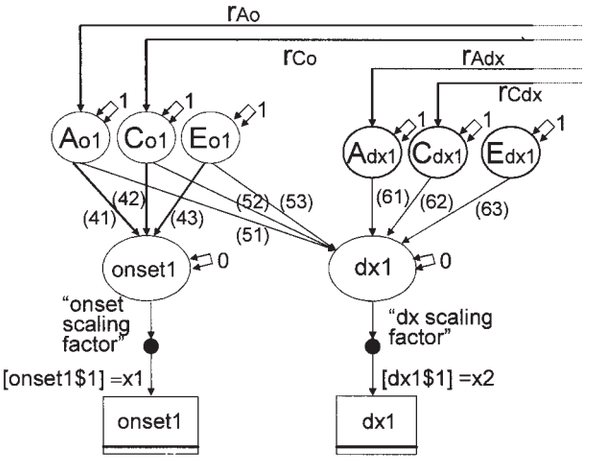

Fig. 7. Mplus specification for a bivariate Cholesky model for two binary variables (Example 7) using the THETA parameterization. The model is shown for one individual (twin 1). Use of the THETA parameterization allows identification of both loadings on E_O (parameters 43 and 53). t1 and t2 are fixed values estimated in a prior run.

model (Example 5). Although the WLSMV values cannot be used for exact comparisons of model fits (see the Mplus User's Guide for details), the large difference in fit suggests the mediation hypothesis can be rejected for this sample. The basis for this can be seen by examining the parameter values. In Example 5, the dx-onset covariance is attributed almost entirely to genetic sources, whereas the mediation model requires the covariance to be in the same proportions as the reliable variance in onset (i.e., $a_o^2 : c_o^2 : e_o^2 = 44 : 11 : 05$).

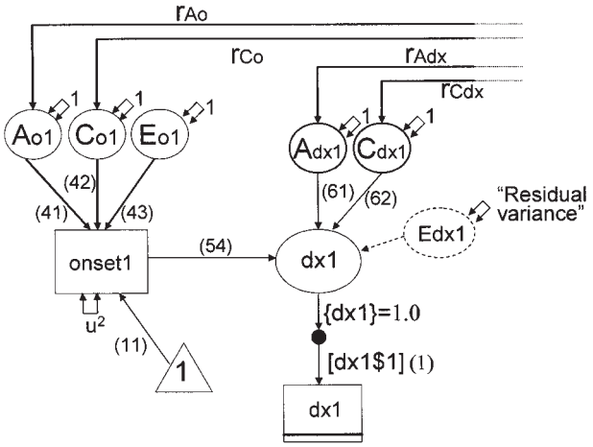

Fig. 6. Mplus specification for a bivariate mediation model to test the hypothesis that onset directly mediates liability to diagnosis (Example 6). The model is shown for one individual (twin 1). Covariation between diagnosis liability and onset is modeled as occurring *through* onset. All sources of covariation (additive genetic, common environment and specific environment) in diagnosis liability occur in the same proportions as the sources of variation in the reliable portion of onset. A fixed estimate of onset unreliability (u^2) is obtained from an external source.

Example 7: Estimating Sources of Covariation for Two Binary Variables

This example uses a standard bivariate model applied to two binary variables and illustrates the usefulness of the THETA parameterization. By fixing the residual variances to 0, all the specific environmental loadings can be estimated. Using the default, DELTA, parameterization only the specific environmental *covariance* is identified; it cannot be separated into two loadings.

The input script used for this example is shown in Table VII. The data employed are from male twin pairs to provide a comparison to the results for females in Examples 5 and 6. The same variables were used, except that onset was recoded into a categorical variable based on a cutting score of age 15. Coding onset as binary provides one way to account for the fact that it is a censored variable. All individuals, including abstainers, can be assigned to a binary variable based on whether they reported drinking by age 15 or not. In the data file, abstainers have been assigned a value of "98" for onset whereas drinkers with missing onset age have

a value of "99." This enables cases with missing data to be excluded from the analysis based on the MISSING statement, but censored cases to be included by recoding the 98 values. Onset is recoded to be 0 if onset ≤ 15 and $= 1$ if onset > 15 (including values of 98). The third DEFINE statement reverses this coding (i.e., recoding onset as 1 for drinking by age 15 and 0 if not), so that early drinking will be positively correlated with alcoholism liability.

In other respects the input script is quite similar to one that would be used for a bivariate model for continuous variables. The exceptions are that the ANALYSIS statement specifies PARAM = THETA and rather than estimating means, there are thresholds fixed to the value estimated in a prior run. Unlike the DELTA parameterization, the residual variances must be fixed to 0. Note that this must be done in a separate MODEL statement for each group listed in the GROUP statement, including the first group. An input script for calculating thresholds is available on the website. Running this yielded the correct values for the thresholds, of .249 for onset and .367 for dx. These were then input into the script using the THETA parameterization.

Selected output is shown in Appendix 7. The proportions of variance are obtained by squaring the values in the StdYX column. Variation in onset is estimated as $a_o^2 = (.619)^2 = .383$, $c_o^2 = (.462)^2 = .213$, $e_o^2 = (.635)^2 = .403$. The proportions of dx shared with onset are $a_{do}^2 = (.546)^2 = .298$, $c_{do}^2 = (-.001)^2 = .000$, $e_{do}^2 = (.039)^2 = .002$, summing to a 30% overlap, virtually all of which is from additive genetic variation contributing to both variables. The residual variation in diagnosis is estimated as $a_d^2 = (.405)^2 = .164$, $c_d^2 = (.293)^2 = .086$, $e_d^2 = (.671)^2 = .450$. In summary, among males, 55% of the variation in liability for alcoholism is estimated as familial, the majority of this is due to genetic variation, including 30% overlapping with onset.

DISCUSSION

The examples illustrate the usefulness of Mplus for structural modeling analyses with data from twins. Mplus has several advantages compared to most other available software programs, including the ability to analyze a combination of categorical and continuously scored dependent variables, to use raw data input, and to obtain rapid convergence of models using multivariate categorical data. Other useful features are the ability to recode variables within the context of the model script, to select subsets of cases based on group membership or values of variables in the dataset, and to include cases with incomplete data.

The program defaults can be very useful, allowing specification of complex, multiple-group models in brief scripts. However (as with other structural modeling software programs), the defaults can also lead to unintended consequences. When developing new models, it is useful to study the specification section of the output (obtained by requesting TECH1) to ensure that all the variances and covariances one believes are fixed to 1 or 0 are correctly specified.

A major disadvantage of Mplus for behavior genetic research is that the program does not have user-defined boundary or other nonlinear constraints. This may limit the application of Mplus for advanced models, such as for assortative mating. When using multivariate models or testing for sex differences, the user may wish to ensure that all loadings are estimated as non-negative. As shown in Appendix 8 and Fig. 8, the program input can be rewritten to keep all loadings non-negative by estimating the square roots of the loadings. However, this is cumbersome with complex models. A related issue is that multiple models may be required to test whether a parameter falls outside its theoretical boundaries (e.g., a male-female genetic correlation <0 or $>.5$), and if so, to constrain it to be inside of the boundary area. This is an imperfect solution as other modifications to the model may alter the estimate of the now-fixed parameters.

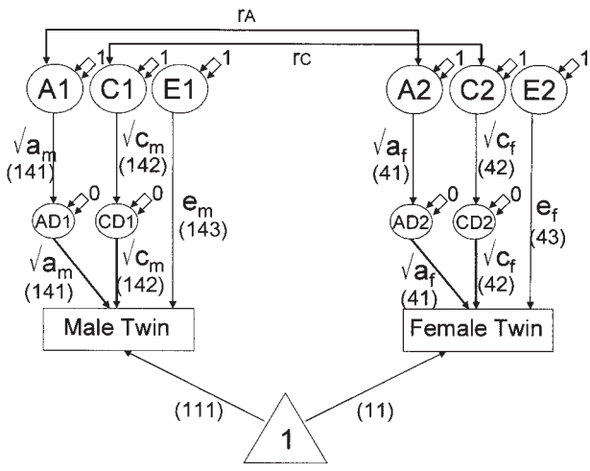

Fig. 8. Mplus specification for a nonscalar sex differences model, parameterized to ensure male-female covariances are non-negative. Parameter labels correspond to the input script in Appendix 8. The parameters as shown apply to opposite-sex pairs. This model uses placeholders (AD_1, CD_1, AD_2, CD_2) to estimate the square roots of the additive genetic and common environment loadings.

Another disadvantage is that (like most other structural modeling programs) the currently available version of Mplus is not able to employ incomplete data with categorical dependent variables. (The next release of Mplus, version 3, will include ML estimation for models with categorical outcomes and will also handle missing data assuming MAR [B. Muthén, personal communication.]) It is possible to separate the observations into groups based on the pattern of missing data, create "pseudo" data for the groups with missing information, and use filter matrices and invariance constraints over groups to have only the true data contribute to the model estimates (see McArdle and Prescott [1996] for an continuous variable example using LISREL). However, this becomes unwieldy when there are many patterns of missing data and is correct only for OLS estimation because when using WLS, inclusion of the pseudo-data will alter the weight matrices.

Some other minor issues associated with the current version of Mplus can be frustrating if the user is unaware of them. When fitting multiple group models to multi-category variables, the first group listed in the Group statement must have data at all levels of the variable or problems will be encountered when reading in the data file. The MISSING value statement allows a range of values to be specified using a dash (e.g., `MISSING=ALL(95-99);`) but care should be used to specify this correctly if the input file uses a negative number as a missing value code.

The examples presented in this paper represent just a few of the models that have been applied to twin and family data. Other advanced applications have been implemented using other programs (e.g., Neale and Cardon, 1992), many of which are possible in Mplus. There are several features of Mplus that were not illustrated here but may be of use to behavior geneticists, including exploratory factor analysis, multilevel modeling, latent variable mixture models, and Monte Carlo simulation studies.

It is worth noting that there are other structural modeling programs that estimate models based on multiple groups (e.g., AMOS, Arbuckle and Wothke, 1999; EQS, Bentler, 1989), but have not been widely employed for behavior genetic applications. Categorical data from twins have also been analyzed using programs developed for pedigree data (e.g., Hannah *et al.,* 1983), and logit analysis (e.g., Kaprio *et al.,* 1981). Other programs for multivariate data analysis such as SAS PROC MIXED (SAS Institute, 1999) have been adapted to estimate sources of covariance among family members (e.g., Guo and Wang, 2002; Prescott and Kendler, 2001). There is no doubt that other programming options and analytic approaches will appear in the future. Criteria to consider when comparing alternative programs include the ease of writing program input, interpretation of program output, options for numerical optimization, speed of computation, calculation of standard errors, availability of user-defined constraints, user support, and the costs of software and training.

The advantages of using Mplus relative to other available software programs will depend on the specific application and the features valued by the researcher. For twin and family researchers experienced with other programs, the main advantages of Mplus are the ability to analyze a combination of continuous and categorical dependent variables, rapid computational speed, and ease of handling common forms of incomplete data. For researchers new to behavior genetic analyses and those unfamiliar with matrix algebra, the simplicity and flexibility of Mplus programming make it an alternative worth considering.

APPENDIX 1

Selected Output from a Univariate Twin Model for a Continuous Variable (Example 1)

```
SUMMARY OF ANALYSIS
Number of groups                                         2
Number of observations
   Group MZ                                           1000
   Group DZ                                           1000
Number of y-variables                                    2
Number of x-variables                                    0
Number of continuous latent variables                    6
```

(*Continued on next page*)

APPENDIX 1

Selected Output from a Univariate Twin Model for a Continuous Variable (Example 1)
(*Continued from previous page*)

```
Observed variables in the analysis
   Y1       Y2
   Grouping variable      ZYG
Continuous latent variables in the analysis
   A1        A2           C1          C2            E1         E2
Estimator                                                    ML
Maximum number of iterations                               1000
Convergence criterion                                  0.500D-04
Maximum number of steepest descent iterations                20

Input data file(s)  example1.dat
Input data format  FREE
SAMPLE STATISTICS
    SAMPLE STATISTICS FOR MZ
         Means
            Y1            Y2
     1      100.385       100.235
         Covariances
            Y1            Y2
 Y1          98.756
 Y2          57.743        96.296
         Correlations
            Y1            Y2
 Y1           1.000
 Y2           0.592         1.000
    SAMPLE STATISTICS FOR DZ
         Means
            Y1            Y2
     1      100.116       100.076
         Covariances
            Y1            Y2
 Y1          90.747
 Y2          37.176       101.818
         Correlations
            Y1            Y2
 Y1           1.000
 Y2           0.387        1.000

THE MODEL ESTIMATION TERMINATED NORMALLY
TESTS OF MODEL FIT
Chi-Square Test of Model Fit
          Value                                          4.818
          Degrees of Freedom                             6
          P-Value                                        0.5673
...
Loglikelihood
          H0 Value                                  -14524.953
          H1 Value                                  -14522.545
...
```

APPENDIX 1

Selected Output from a Univariate Twin Model for a Continuous Variable (Example 1) (*Concluded*)

```
MODEL RESULTS
                                Estimates      S.E.      Est./S.E.        Std        StdYX
Group MZ
 A1    BY     Y1                 6.255       0.481       12.999         6.255       0.636
 A2    BY     Y2                 6.255       0.481       12.999         6.255       0.636
 C1    BY     Y1                 4.239       0.621        6.822         4.239       0.431
 C2    BY     Y2                 4.239       0.621        6.822         4.239       0.431
 E1    BY     Y1                 6.298       0.138       45.493         6.298       0.640
 E2    BY     Y2                 6.298       0.138       45.493         6.298       0.640

 A1    WITH   A2                 1.000       0.000        0.000         1.000       1.000
              C1                 0.000       0.000        0.000         0.000       0.000
              C2                 0.000       0.000        0.000         0.000       0.000
              E1                 0.000       0.000        0.000         0.000       0.000
              E2                 0.000       0.000        0.000         0.000       0.000

 A2    WITH   C1                 0.000       0.000        0.000         0.000       0.000
              C2                 0.000       0.000        0.000         0.000       0.000
              E1                 0.000       0.000        0.000         0.000       0.000
              E2                 0.000       0.000        0.000         0.000       0.000

 C1    WITH   C2                 1.000       0.000        0.000         1.000       1.000
              E1                 0.000       0.000        0.000         0.000       0.000
              E2                 0.000       0.000        0.000         0.000       0.000

 C2    WITH   E1                 0.000       0.000        0.000         0.000       0.000
              E2                 0.000       0.000        0.000         0.000       0.000

 E1    WITH   E2                 0.000       0.000        0.000         0.000       0.000
Means
    A1                          0.000       0.000        0.000         0.000       0.000
    A2                          0.000       0.000        0.000         0.000       0.000
    C1                          0.000       0.000        0.000         0.000       0.000
    C2                          0.000       0.000        0.000         0.000       0.000
    E1                          0.000       0.000        0.000         0.000       0.000
    E2                          0.000       0.000        0.000         0.000       0.000
Intercepts
    Y1                        100.196       0.189      529.142       100.196      10.185
    Y2                        100.196       0.189      529.142       100.196      10.185
Group DZ
 A1    BY     Y1                 6.255       0.481       12.999         6.255       0.636
 A2    BY     Y2                 6.255       0.481       12.999         6.255       0.636
 C1    BY     Y1                 4.239       0.621        6.822         4.239       0.431
 C2    BY     Y2                 4.239       0.621        6.822         4.239       0.431
 E1    BY     Y1                 6.298       0.138       45.493         6.298       0.640
 E2    BY     Y2                 6.298       0.138       45.493         6.298       0.640

 A1    WITH   A2                 0.500       0.000        0.000         0.500       0.500
              C1                 0.000       0.000        0.000         0.000       0.000
              C2                 0.000       0.000        0.000         0.000       0.000
              E1                 0.000       0.000        0.000         0.000       0.000
              E2                 0.000       0.000        0.000         0.000       0.000

 A2    WITH   C1                 0.000       0.000        0.000         0.000       0.000
              C2                 0.000       0.000        0.000         0.000       0.000
              E1                 0.000       0.000        0.000         0.000       0.000
              E2                 0.000       0.000        0.000         0.000       0.000

 C1    WITH   C2                 1.000       0.000        0.000         1.000       1.000
...
```

... indicates lines omitted

APPENDIX 2

Selected Output from a Univariate-Twin Model for a Binary Variable (Example 2)

```
SAMPLE STATISTICS
      ESTIMATED SAMPLE STATISTICS FOR MZ
             SAMPLE THRESHOLDS
                Y1$1           Y2$1
       1          0.015       -0.010
             SAMPLE TETRACHORIC CORRELATIONS
                Y1             Y2
 Y1
 Y2               0.583
      ESTIMATED SAMPLE STATISTICS FOR DZ
             SAMPLE THRESHOLDS
                Y1$1           Y2$1
       1           0.040             0.023
             SAMPLE TETRACHORIC CORRELATIONS
               Y1             Y2
 Y1
 Y2                0.405
...
MODEL RESULTS
                          Estimates      S.E.      Est./S.E.       Std       StdYX
Group MZ
 A1       BY     Y1         0.596       0.096       6.193       0.596      0.596
 A2       BY     Y2         0.596       0.096       6.193       0.596      0.596
 C1       BY     Y1         0.477       0.100       4.788       0.477      0.477
 C2       BY     Y2         0.477       0.100       4.788       0.477      0.477

 A1       WITH   C1         0.000       0.000       0.000       0.000      0.000
                 C2         0.000       0.000       0.000       0.000      0.000
                 A2         1.000       0.000       0.000       1.000      1.000

 A2       WITH   C1         0.000       0.000       0.000       0.000      0.000
                 C2         0.000       0.000       0.000       0.000      0.000

 C1       WITH   C2         1.000       0.000       0.000       1.000      1.000

 Means
    A1                      0.000       0.000       0.000       0.000      0.000
    A2                      0.000       0.000       0.000       0.000      0.000
    C1                      0.000       0.000       0.000       0.000      0.000
    C2                      0.000       0.000       0.000       0.000      0.000
 Thresholds
    Y1$1                    0.017       0.023       0.739       0.017      0.017
    Y2$1                    0.017       0.023       0.739       0.017      0.017
 Variances
    A1                      1.000       0.000       0.000       1.000      1.000
    A2                      1.000       0.000       0.000       1.000      1.000
    C1                      1.000       0.000       0.000       1.000      1.000
    C2                      1.000       0.000       0.000       1.000      1.000
 Scales
    Y1                      1.000       0.000       0.000       1.000      1.000
    Y2                      1.000       0.000       0.000       1.000      1.000
Group DZ
 A1     BY       Y1         0.596       0.096       6.193       0.596      0.596
 A2     BY       Y2         0.596       0.096       6.193       0.596      0.596
 C1     BY       Y1         0.477       0.100       4.788       0.477      0.477
 C2     BY       Y2         0.477       0.100       4.788       0.477      0.477
 A1     WITH     C1         0.000       0.000       0.000       0.000      0.000
                 C2         0.000       0.000       0.000       0.000      0.000
                 A2         1.000       0.000       0.000       1.000      1.000
...
R-SQUARE
Group MZ
    Observed  Residual
    Variable  Variance  R-Square
    Y1          0.417     0.583
    Y2          0.417     0.583
Group DZ
    Observed  Residual
    Variable  Variance  R-Square
    Y1          0.417     0.583
    Y2          0.417     0.583
```

APPENDIX 3

Summary of Mplus Input for Continuous and Categorical Variables

	VARIABLE TYPE & PARAMETERIZATION		
	Continuous	Binary − DELTA	Binary − THETA
TITLE	Two-Group ACE Model		
DATA	FILE = example.dat;		
VARIABLE	NAMES = famno zyg y1 y2; USEVAR = y1 y2; GROUP = zyg(1=MZ 2=DZ);		
		CATEGORICAL = y1 y2;	
ANALYSIS	TYPE = MEANSTRUCTURE;		
		PARAM=DELTA;[a]	PARAM=THETA;
MODEL	[y1](1); [y2] (1);	[y1$1] (1); [y2$1] (1);	[y1$1@**x**]; [y2$1@**x**];[b]
	y1@0; y2@0;	{y1@1 y2@1};	
	A1 by y1 (11); A2 by y2 (11); C1 by yl (12); C2 by y2 (12); E1 by y1 (13); E2 by y2 (13);	A1 by y1 (11); A2 by y2 (11); C1 by y1 (12); C2 by y2 (12);	*same as continuous*
	[A1-E2@0]; A1-E2@1; A1-A2 with C1-E2@0; C1-C2 with E1-E2@0; A1 with A2@1; C1 with C2@1; El with E2@0;	[A1-C2@0]; A1-C2@1; A1-A2 with C1-C2@0; A1 with A2@1; C1 with C2@1;	*same as continuous*
MODEL MZ			y1@0; y2@0;
MODEL DZ	A1 with A2@0.5;	A1 with A2@0.5;	A1 with A2@0.5; yl@0; y2@0;
OUTPUT	SAMP RES STAND;		
Obtaining parameter estimates[c]	mean = par 1 a = par 11 c = par 12 e = par 13	threshold = par 1 a = par 11 c = par 12 e^2 = residual variance	threshold = **x** * scaling factor a= par11 * scaling factor c= par12 * scaling factor e= par13 * scaling factor

[a] Mplus default, statement not required
[b] **x** = any positive number; if **x** is fixed at the true threshold value (e.g., obtained in a prior run using the Delta parameterization), the scaling factor = 1 and t-values will be accurate
[c] par = parameter; residual variance and scaling factor are printed when output RES is requested
Note: Blank cell indicates no statement required; starting values for estimated parameters are not shown, but may be needed

ACKNOWLEDGMENTS

Preparation of this manuscript was supported by K01-AA-00236. Data collection was supported by National Institutes of Health Grants R01-MH/AA-49492 and R01-AA/DA-09095. The author is grateful to Jack McArdle, Lindon Eaves, Matt McGue, Bengt Muthén, Mike Neale, Wendy Slutske, Jonathan Kuhn, Eric Turkheimer, and an anonymous reviewer for helpful comments on earlier versions of this article. The author has no financial interest in any of the software discussed.

REFERENCES

American Psychiatric Association. (1994). *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.). Washington, DC: American Psychiatric Association.

Arbuckle, P., and Wothke, W. (1999). *AMOS User's Guide, version 4.* Chicago, IL: SPSS.

Behrman, J. R., Taubman, P., and Wales, T. J. (1978). The roles of genetics and environment in the distribution of earnings. In Griliches, Z., Krelle, W., Krupp, H.-J., and Kyn O. (eds.), *Income Distribution and Economic Inequality* (pp. 220–239). New York: J. Wiley and Sons.

Bentler, P. M. (1989). *EQS: Structural Equations Program Manual.* Los Angeles, CA: BMDP Statistical Software.

Boomsma, D. I., and Molenaar, P. C. (1986). Using LISREL to analyze genetic and environmental covariance structure. *Behav. Genet.* **16**:237–250.

Cantor, R. M., and Nance, W. E. (1981). An application of confirmatory factor analysis to individual finger ridge count data from the offspring of MZ twins. *Am. J. Hum. Genet.* **33**:134A.

Cantor, R. M. (1983). A multivariate analysis of ridge count data from the offspring of monozygotic twins. *Acta Genet. Med. Gemellol* **32**:161–207.

Cattell, R. B. (1960). The multiple abstract variance analysis equations and solutions: For nature-nurture research on continuous variables. *Psych. Rev.* **67**:353–377.

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.

Edwards, J. H. (1969). Familial predisposition in man. *Br. Med. Bull.* **25**:58–64.

Falconer, D. S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann. Hum. Genet.* **29**:51–76.

Fulker, D. W. (1973). A biometrical genetic approach to intelligence and schizophrenia. *Soc. Biol.* **20**:266–275.

Fulker, D. W. (1978). Multivariate extensions of a biometrical model of twin data. *Prog. Clin. Biol. Res.* **24A**:217–236.

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

Hannah, M. C., Hopper, J. L., and Mathews, J. D. (1983). Twin concordance for a binary trait. I. Statistical models illustrated with data on drinking status. *Acta Genet. Med. Gemellol* **32**:127–137.

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

Joreskog, K. G. (1973). Analysis of covariance structures. In Krishnarah, P. R. (ed.), *Multivariate Analysis III: Proceedings of the Third International Symposium on Multivariate Analysis.* New York: Academic Press.

Joreskog, K. G., and Sorbom, D. (1977). *LISREL: Analysis of Linear Structural Relationships by the Method of Maximum Likelihood.* Chicago, IL: National Educational Resources.

Joreskog, K. G., and Sorbom, D. (1986). *PRELIS: A Preprocessor for LISREL.* Mooresville, IN: Scientific Software.

Kaprio, J., Sarna, S., and Koskenvuo, M. (1981). Multivariate logit analysis of concordance ratios for qualitative traits in twin studies. *Acta Genet. Med. Gemellol.* **30**:267–274.

Kendler, K. S., and Prescott, C. A. (1999). A population-based twin study of lifetime major depression in men and women. *Arch. Gen. Psychiatry* **56**:39–44.

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

Mather, K., and Jinks, J. L. (1971). *Biometrical Genetics* (2nd ed.). London: Chapman and Hall.

McArdle, J. J., Connell, J. P., and Goldsmith, H. H. (1980). Structural modeling of stability and genetic influences: Some results from a longitudinal study of behavioral style. *Behav. Genet.* **10**:487.

McArdle, J. J., and Goldsmith, H. H. (1990). Alternative common factor models for multivariate biometric analyses. *Behav. Genet.* **20**:569–608.

McArdle, J. J., and McDonald, R. P. (1984). Some algebraic properties of the reticular action model. *Br. J. Math. Stat. Psych.* **37**:234–251.

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

Meyer, J. M., Eaves, L. J., Heath, A. C., and Martin, N. G. (1991). Estimating genetic influences on the age-at-menarche: A survival analysis approach. *Am. J. Med. Genet.* **39**:148–154.

Muthén, B. O. (1988). *LISCOMP: Analysis of Linear Structural Equations Using a Comprehensive Measurement Model* (2nd ed.). Mooresville, IN: Scientific Software.

Muthén, B., and Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus, version 5, December 9, 2002. www.statmodel.com/mplus/examples/webnote.html

Muthén, B., Shedden, K., and Spisic, D. (1999). *General Latent Variable Mixture Modeling.* Tech. Report. Santa Monica, CA: Muthén and Muthén.

Muthén, L. K., and Muthén, B. O. (1998). *Mplus User's Guide.* Los Angeles, CA: Muthén and Muthén.

Muthén, L. K., and Muthén, B. O. (2002). *Mplus VERSION 2.12: Addendum to the Mplus user's Guide.* Los Angeles, CA: Muthén and Muthén.

Neale, M. C. (1991). *Mx: Statistical Modeling,* Department of Human Genetics. Richmond, VA: Virginia Commonwealth University.

Neale, M. C., and Cardon, L. R. (1992). *Methodology for Genetic studies of Twins and Families.* Dordrecht: Kluwer Academic Publishers.

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

Prescott, C. A. (1991). *Clinical, Psychometric and Biometric Assessment of Schizophrenia: A Psychiatric Twin Study.* University of Virginia. Unpublished Dissertation.

Prescott, C. A., Aggen, S. H., and Kendler, K. S. (1999). Sex differences in the sources of genetic liability to alcohol abuse and dependence in a population based sample of U.S. twins. *Alcohol Clin. Exp. Res.* **23**:1136–1144.

Prescott, C. A., Cross, R. J., Kuhn, J. W., Horn, J. L., and Kendler, K. S. (in press). Is risk for alcoholism mediated by individual differences in drinking motivations? *Alcohol Clin. Exp. Res.*

Prescott, C. A., and Gottesman, I. I. (1990). Biometric prediction of genetic liability for schizophrenia. *Behav. Genet.* **20**:742.

Prescott, C. A., and Kendler, K. S. (1999a). Age at first drink and risk for alcoholism: A noncausal association. *Alcohol Clin. Exp. Res.* **23**:101–107.

Prescott, C. A., and Kendler, K. S. (1999b). Genetic and environmental contributions to alcohol abuse and dependence in a population-based sample of male twins. *Am. J. Psychiatry* **156**:34–40.

Prescott, C. A., and Kendler, K. S. (2001). Associations between marital status and alcohol consumption in a longitudinal study of female twins. *J. Stud. Alcohol* **62**:589–604.

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

Smith, C. (1974). Concordance in twins: Methods and interpretation. *Am. J. Hum. Genet.* **6**:454–466.

Waller, N. G., and Muthén, B. O. (1992). Genetic Tobit factor analysis: Quantitative genetic modeling with censored data. *Behav. Genet.* **22**:265–292.

Yashin, A. I., Iachine, I. A., and Harris, J. R. (1999). Half of the variation in susceptibility to mortality is genetic: Findings from Swedish twin survival data. *Behav. Genet.* **29**:11–19.

Edited by Wendy Slutske