Title

On Fitting a Multivariate Two-Part Latent Growth Model

Journal

Structural Equation Modeling: A Multidisciplinary Journal, 21(1)

ISSN

1070-5511 1532-8007

Authors

Xu, Shu
Blozis, Shelley A
Vandewater, Elizabeth A

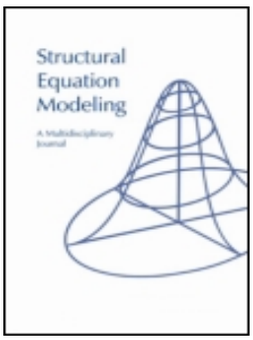
Publication Date

2014-01-31

DOI

10.1080/10705511.2014.856699

Peer reviewed



Routledge
Taylor & Francis Group

Structural Equation Modeling: A Multidisciplinary Journal

ISSN: 1070-5511 (Print) 1532-8007 (Online) Journal homepage: http://www.tandfonline.com/loi/hsem20

On Fitting a Multivariate Two-Part Latent Growth Model

Shu Xu , Shelley A. Blozis & Elizabeth A. Vandewater

To cite this article: Shu Xu , Shelley A. Blozis & Elizabeth A. Vandewater (2014) On Fitting a Multivariate Two-Part Latent Growth Model, Structural Equation Modeling: A Multidisciplinary Journal, 21:1, 131-148, DOI: 10.1080/10705511.2014.856699

To link to this article: http://dx.doi.org/10.1080/10705511.2014.856699

Structural Equation Modeling: A Multidisciplinary Journal, 21: 131–148, 2014

ISSN: 1070-5511 print / 1532-8007 online
DOI: 10.1080/10705511.2014.856699

Routledge
Taylor & Francis Group

TEACHER'S CORNER

On Fitting a Multivariate Two-Part Latent Growth Model

Shu Xu,[1] Shelley A. Blozis,[2] and Elizabeth A. Vandewater[3]
[1]*New York University*
[2]*University of California, Davis*
[3]*The University of Texas at Austin*

A 2-part latent growth model can be used to analyze semicontinuous data to simultaneously study change in the probability that an individual engages in a behavior, and if engaged, change in the behavior. This article uses a Monte Carlo (MC) integration algorithm to study the interrelationships between the growth factors of 2 variables measured longitudinally where each variable can follow a 2-part latent growth model. A SAS macro implementing M*plus* is developed to estimate the model to take into account the sampling uncertainty of this simulation-based computational approach. A sample of time-use data is used to show how maximum likelihood estimates can be obtained using a rectangular numerical integration method and an MC integration method.

Keywords: longitudinal semicontinuous variables, Monte Carlo integration, multivariate two-part latent growth curve model

A common aim within the behavioral sciences is understanding change in a measured response. Longitudinal data allow for assessment of change. With longitudinal data, individuals are observed at multiple time points to gain an understanding of the ways in which a behavior changes according to time. The appropriate statistical method used for the analysis of longitudinal data is based on theoretical considerations concerning the behavior under study, as well as characteristics of the data, including the response distribution. Measures of substance use (e.g., Blozis, Feldman, & Conger, 2007; Witkiewitz & Masyn, 2008), problem behaviors (e.g., Petras, Nieuwbeerta, & Piquero, 2010; Vazsonyi & Keiley, 2007), and time use are examples of variables that often follow a semicontinuous distribution (Olsen & Schafer, 2001) in which a variable takes on a high proportion of zeros with the remaining scores being positive and continuous. In considering the distribution of time spent using a computer among adolescents, for example, a large proportion of zeros results if a substantial portion of a sample report that they did not use a computer.

In comparison to continuous data, the semicontinuous data just described may yield two distinct pieces of information about a behavioral response. The first describes whether or not an individual engaged in a behavior, as indicated by a response value of 0 versus any positive value. The second is the magnitude of the response when it does occur, as indicated by positive response values. Longitudinal semicontinuous data present an analytic challenge given both the response distribution and the need to model change or growth in the response. This is a challenge because responses with this distribution are difficult to classify into any general distribution commonly considered for latent growth models, including linear and generalized linear models.

Different approaches to the analysis of semicontinuous data have been considered. One approach is to apply methods that are appropriate for normally distributed data. In a study of drug use, for instance, given the high frequency of reporting a zero in a single type of drug use, a composite item could be created as a sum of multiple drug uses or a sum of drug

Correspondence should be addressed to Shelley A. Blozis, Department of Psychology, University of California, Davis, One Shields Avenue, Davis, CA 95616, USA. E-mail: sablozis@ucdavis.edu

use in various occasions to reduce the skewness of the data (e.g., Hoffmann, Cerbone, & Su, 2000). This approach might help to satisfy the analytical model assumptions; however, researchers might not be able to directly address questions that are closely related to the particular outcomes of interest. Thus, inference is limited to the composite item. Another approach is to dichotomize the response into two discrete categories with one that denotes an absence of the behavior (i.e., those responses equal to zero) and the other that denotes that the behavior was observed (i.e., positive scores are set equal to a value different from 0, such as 1; e.g., Duncan, Duncan, & Strycker, 2006). A consequence of this procedure, however, is that individual differences in the positive responses are likely to be eliminated. That is, the analysis ignores the extent to which individuals vary in terms of the magnitude of their responses.

In some cases it might be important to make the distinction between the two aspects of such behaviors: whether or not a behavior occurred and the magnitude of the response when the behavior did occur. In this situation, an analysis of a semicontinuous variable could best be carried out with a method that simultaneously addresses these two features. In separating the two features, it might also be possible to study different predictors of each aspect of a behavior, rather than to assume that the same predictors relate to each. Specifically, one set of predictors might account for the probability that a behavior occurs, and the same or a different set might best account for the magnitude of the response when it does occur. These issues could be addressed by a two-part model.

A two-part latent growth model was developed for semicontinuous variables observed over time while accounting for the within-individual correlations typically observed in longitudinal data (Olsen & Schafer, 2001; Tooze, Grunwald, & Jones, 2002). In a two-part latent growth model, two new variables are created to represent (a) the presence or absence of a behavior, and (b) the magnitude of the behavior when it occurs. Two submodels, a random-effects logistic growth model and a latent growth model for normally distributed data, are used to model the binary and the positive responses, respectively. The random effects of each submodel may covary at the second level of the model. Thus, the two model components are jointly estimated. Allowing the random effects of the two model components to covary might be important for model estimation; more specifically, an incorrect assumption of independence between the random effects of the two submodels could result in biased estimates (Su, Tom, & Farewell, 2009). In a two-part growth curve model with random intercepts in both the binary and continuous submodels, for example, a failure to model a positive correlation between the random intercepts might result in a positive bias in the estimate of the intercept of the continuous submodel. A two-part latent growth model has been considered for longitudinal substance-use data (Blozis et al., 2007; Brown, Catalano, Fleming, Haggerty, & Abbott, 2005; Liu, Ma, & Johnson, 2008; Weaver, Cheong, MacKinnon, & Pentz, 2011).

In many longitudinal studies, two or more semicontinuous variables are measured across time with the goal of understanding how different variables might be related to one another over time. Time-use data, for example, have been used to study possible links between electronic media and other behaviors (e.g., reading), a topic of interest to policymakers and researchers (Altheide, 1997). Researchers might be interested in questions such as these: How much time do adolescents spend using a computer and reading? Does time spent in these behaviors change as children grow older? What is the relationship between using a computer and reading over time?

To study the relationships between longitudinal measures, one variable could be hypothesized, for instance, to affect another variable within each measurement occasion (at the occasional level, or Level 1), or change in one variable might be hypothesized to be related to change in another variable (at the individual level, or Level 2). The latter kind of model is referred to as a multivariate latent growth model (Blozis, 2004, 2007; MacCallum, Kim, Malarkey, & Kiecolt-Glaser, 1997; Reinsel, 1992). In a multivariate latent growth model, two or more normal variables are each assumed to follow a latent growth model. At the second level of the model for observed measures, for example, the random effects that characterize growth in each response can covary. In this way, the model allows for the study of the linear associations between features of changes, such as intercepts and slopes, of multiple variables.

Although the concept of a multivariate two-part latent growth model is a natural extension of the model for a single semicontinuous response, estimation of the model can be computationally intensive and difficult. The joint distribution of a binary and continuous variable is in general not easy to compute, particularly when the variables are observed longitudinally (Skrondal & Rabe-Hesketh, 2004; Verbeke & Davidian, 2008). Thus, the joint modeling of two or more semicontinuous variables consequently increases the computational load. These issues suggest a great challenge in the estimation of a multivariate model for semicontinuous data.

Other approaches have been proposed for the analysis of positive continuous data with a high proportion of zeros. These methods can be classified as either one-part or two-part (joint) models. One-part models (e.g., a latent growth model or a generalized latent growth model) were described earlier in this article. A two-part latent growth model is considered appropriate if the zeros in a data set represent true zeros (Olsen & Schafer, 2001), such that the zero represents the actual level of the outcome. A competing model, a Tobit latent growth model (Wang, Zhang, McArdle, & Salthouse, 2008) based on a Tobit model (Tobin, 1958) for cross-sectional censored data, is considered appropriate if the zeros result from a censoring process. In this case, zero could be a proxy of a missing or negative value. In a more

complicated situation where samples are drawn from more than one population, a mixture model approach (e.g., a finite mixture model) could be used (Muthén, 2001). In other cases, if the outcome variable is a count variable with a large proportion of zeros, then a zero-inflated Poisson or a zero-inflated negative binomial model (Hall, 2000) might be most appropriate.

This article considers a multivariate two-part latent growth model to simultaneously investigate two longitudinal behavioral outcomes, each of which is assumed to follow a two-part latent growth model. This allows one to study the relationships between latent characteristics of change in each of the two aspects of two different behaviors. This represents a complex joint modeling of two continuous and two dichotomous responses. In the reminder of this article, we first provide a brief review of a two-part latent growth model. We then develop a multivariate two-part latent growth model. We discuss assumptions about missing data and describe available estimators and computational algorithms. The statistical software package M*plus* (Muthén & Muthén, 1998–2010) can be used to obtain maximum likelihood (ML) estimates of such complex models using numerical integration methods. A SAS macro is developed to invoke M*plus* to obtain ML estimates with standard errors that are approximated by first-order derivatives (denoted henceforth as MLF). The macro relies on a Monte Carlo (MC) integration method for a multivariate two-part latent growth model that joins two semicontinuous measures. An empirical example from a time-use study illustrates an application of a multivariate two-part growth model. Related issues are then discussed.

A TWO-PART LATENT GROWTH MODEL

Readers are referred to Olsen and Schafer (2001) and Tooze et al. (2002) for details and applications of a two-part latent growth model. The model is briefly reviewed here. Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})'$ for individual i be a set of n_i repeated semicontinuous responses measured at occasion $j = 1, 2, \ldots, n_i$, where the subscript i on n_i indicates that individuals can be observed a different number of times. The repeated measures are assumed to be observed according to time, denoted by $\mathbf{t_i} = (t_{i1}, \ldots, t_{in_i})'$, where the specific times of measurement can vary between individuals. It is assumed for convenience that the values of y could range from zero to any positive number.

Assuming that a portion of the responses are equal to zero and the remaining values are positive, two new variables are created. For individual i, let $\mathbf{u}_i = (u_{i1}, \ldots, u_{in_i})'$ be a set of n_i binary responses with values equal to 1 if the corresponding values of $\mathbf{y}_i$ are greater than zero and equal to 0 otherwise. Let $\mathbf{m}_i = (m_{i1}, \ldots, m_{ik_i})'$ be a set of k_i continuous and positive responses that are equal to the corresponding values of $\mathbf{y}_i$ if those values are greater than zero; values are missing otherwise. A joint response set $\mathbf{y}_i{}^*$ is then created by stacking the two new variables: $\mathbf{y}_i{}^* = (\mathbf{u}_i{}', \mathbf{m}_i{}')'$.

The set of responses $\mathbf{y}_i{}^*$ is assumed to follow a two-part latent growth model. For the binary component $\mathbf{u}_{ij}$, a generalized latent growth model (Liang & Zeger, 1986) that assumes a Bernoulli distribution is applied. Using a logit link function, repeated measures of the log-odds of the binary outcome are considered a linear function of time $\mathbf{t}_i$. Specifically, let π_{ij} denote the probability of observing a behavior (i.e., $P(u_{ij} = 1)$) for individual i, then $\eta_{ij} = \text{logit}(P(u_{ij} = 1)) = \log(\pi_{ij}/(1 - \pi_{ij}))$ is the corresponding log-odds. Assuming linear change in the log-odds, for example, a latent logit growth model for a response at the jth occasion as a function of t_{ij}, is

$$\eta_{ij} = \alpha_{0i} + \alpha_{1i} t_{ij}$$

The coefficients α_{0i} and α_{1i} may each be sums of fixed and random effects, $\alpha_{0i} = \alpha_0 + a_{0i}$ and $\alpha_{1i} = \alpha_1 + a_{1i}$, where the expected values of α_{0i} and α_{1i} are α_0 and α_1, respectively. The coefficients α_0 and α_1 denote the expected log-odds of the behavior for $t_{ij} = 0$ and the expected change in the log-odds per unit of time, respectively. The individual-specific random effects a_{0i} and a_{1i} are assumed to be normal and deviate about their respective fixed effects. The random effects can covary. The variances of a_{0i} and a_{1i} are summary measures of the extent to which individuals vary with regard to these change features.

Next, the set of positive and continuous responses are assumed to follow a latent growth model (Meredith & Tisak, 1984, 1990). Conditional on whether the behavior is observed at time t_{ij}, m_{ij} is considered a function of time t_{ij} plus an error term, ε_{ij}. Assuming the behavior changes at a constant rate, for example,

$$m_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + \varepsilon_{ij},$$

where, for individual i, β_{0i} is the expected response given $t_{ij} = 0$, and β_{1i} is the expected change in the response per unit of time. As in the logit model, the coefficients β_{0i} and β_{1i} may be sums of fixed and random effects. The fixed effects β_0 and β_1 are common to all individuals and describe the population-level response. The random effects b_{0i} and b_{1i} are assumed to be normal and to deviate about their corresponding fixed effects at the second level of the model. The random effects at the second level can covary with each other but are assumed to be independent of the time-specific error ε_{ij} that varies at the first level. The variances of the random effects at the second level summarize the extent of individual differences in the change features. The set of occasion-specific errors $\boldsymbol{\varepsilon}_i = (\boldsymbol{\varepsilon}_{i1}, \ldots, \boldsymbol{\varepsilon}_{ik_i})'$ are assumed to be independent between individuals. Conditional on time and the individual-specific growth function, the errors could be assumed to be normally distributed with mean 0 and to have constant variance σ_ε^2 across measurement occasions. In some

applications, the errors might be allowed to covary between occasions or to have nonconstant variance across time.

The models for the binary and continuous responses can be considered jointly by allowing covariances between the random effects at the individual level of each submodel. Assuming two random effects for each submodel, for instance, the random effects of the logit model, $\mathbf{a}_i = (a_{0i}, a_{1i})'$, and those of the continuous model, $\mathbf{b}_i = (b_{0i}, b_{1i})'$, are assumed to have a joint normal distribution:

$$\begin{pmatrix} a_{0i} \\ a_{1i} \\ b_{0i} \\ b_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \mathbf{\Phi} = \begin{pmatrix} \phi_{a00} & & & \\ \phi_{a1a0} & \phi_{a11} & & \\ \phi_{b0a0} & \phi_{b0a1} & \phi_{b00} & \\ \phi_{b1a0} & \phi_{b1a1} & \phi_{b1b0} & \phi_{b11} \end{pmatrix}\right)$$

where the expected values of the random effects are equal to zero, and the matrix $\mathbf{\Phi}$ is a symmetric covariance matrix.

Missing Data

Similar to many longitudinal studies, missing data are likely to result from studies of semicontinuous longitudinal data but possibly due to multiple sources. One source that is shared with other longitudinal studies is that the behavioral measure might be missing for any occasion, such as if an individual misses an interview or has dropped out of the study. If the probability of missing data is independent of the missing data but possibly related to the observed data, the data are said to be missing at random (MAR; Rubin, 1976). Inference of a latent growth model is considered valid if data are MAR. This is also the case for a two-part latent growth model (Olsen & Schafer, 2001).

Another source of missing data is due to the creation of the new variables that are to be analyzed. Specifically, the measure of the conditional magnitude of the response m is treated as missing if an individual did not engage in the behavior (i.e., if $u = 0$, m is conditionally missing). Under a two-part latent growth model, a value of zero for the semicontinuous variable is a random realization of the extent to which an individual engages in the behavior. Thus, the realization of m is also random. Further, the probability that m is missing is dependent on the value of u. As a result, inference from a two-part latent growth model is considered valid if missing data are MAR assuming that u has been observed.

MULTIVARIATE TWO-PART LATENT GROWTH MODEL

A multivariate two-part latent growth model is developed by jointly modeling two or more longitudinal semicontinuous response variables, each assumed to be due to a two-part latent growth model. For ease of presentation, two longitudinal semicontinuous variables are considered. As described later, the repeated measures could be related at both the occasion and the individual levels.

A multivariate response set is created by stacking the separate sets of responses relating to the binary and continuous variables created from each of the two original semicontinuous variables. First, let $\mathbf{u}_{1i} = (u_{1i1}, \ldots, u_{1in_{1i}})'$ and $\mathbf{u}_{2i} = (u_{2i1}, \ldots, u_{2in_{2i}})'$ for individual i be a set of $n_i = n_{1i} + n_{2i}$ binary responses relating to the first and second semicontinuous variables, respectively. Then, let $\mathbf{m}_{1i} = (m_{1i1}, \ldots, m_{1ik_{1i}})'$ and $\mathbf{m}_{2i} = (m_{2i1}, \ldots, m_{2ik_{2i}})'$ for individual i be a set of $k_i = k_{1i} + k_{2i}$ continuous responses relating to the first and second semicontinuous variables, respectively. Finally, let the set $\mathbf{y}_i^* = (u'_{1i}, m'_{1i}, u'_{2i}, m'_{2i})'$ be the multivariate response set.

At the first level of the multivariate two-part latent growth model, the two measured behaviors might be related through the time-specific errors of the continuous parts of each submodel. Let $\boldsymbol{\varepsilon}_{1i} = (\varepsilon_{1i1}, \ldots, \varepsilon_{1ik_{1i}})'$ and $\boldsymbol{\varepsilon}_{2i} = (\varepsilon_{2i1}, \ldots, \varepsilon_{2ik_{2i}})'$ for individual i be the occasion-level errors of the first and second continuous response variables, $\mathbf{m}_{1i}$ and $\mathbf{m}_{2i}$, respectively. The errors are assumed to be normally distributed with means equal to $\mathbf{0}$ and covariance matrix $\mathbf{\Theta}_i$:

$$\begin{pmatrix} \boldsymbol{\varepsilon}_{1i} \\ \boldsymbol{\varepsilon}_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \mathbf{\Theta}_i = \begin{pmatrix} \mathbf{\Theta}_{1i} & \\ \mathbf{\Theta}_{21i} & \mathbf{\Theta}_{2i} \end{pmatrix}\right)$$

where $\mathbf{\Theta}_i$ is a symmetric covariance matrix that contains the covariance matrices of the two continuous response variables, $\mathbf{\Theta}_{1i}$ and $\mathbf{\Theta}_{2i}$, as well as a nonsymmetric matrix of the covariances between the errors of the two variables, $\mathbf{\Theta}_{21i}$. The matrices could vary between individuals with regard to their dimensions, as determined by the number of observations of each variable for the individual, but typically do not differ otherwise. If, for instance, the errors of the models for both $\mathbf{m}_{1i}$ and $\mathbf{m}_{2i}$ are assumed to have constant variances across time and to covary only within the same occasions, the errors might be assumed to have the following distribution:

$$\begin{pmatrix} \varepsilon_{1i1} \\ \varepsilon_{1i2} \\ \vdots \\ \varepsilon_{1ik_{1i}} \\ \varepsilon_{2i1} \\ \varepsilon_{2i2} \\ \vdots \\ \varepsilon_{2ik_{2i}} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \mathbf{\Theta}_i = \begin{bmatrix} \sigma_1^2 & & & & & & & \\ 0 & \sigma_1^2 & & & & & & \\ \vdots & \vdots & \ddots & & & & & \\ 0 & 0 & \ldots & \sigma_1^2 & & & & \\ \sigma_{21} & 0 & \cdots & 0 & \sigma_2^2 & & & \\ 0 & \sigma_{21} & \ldots & 0 & 0 & \sigma_2^2 & & \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \\ 0 & 0 & \ldots & \sigma_{21} & 0 & 0 & \ldots & \sigma_2^2 \end{bmatrix}\right),$$

where σ_1^2 and σ_2^2 are the variance of the occasion-specific error of the continuous response from the first and second behavior, respectively, and σ_{21} is the covariance between the errors of the two responses within the same occasion.

At the second level of the model, the two measured behaviors could be related through the covariances between the latent change characteristics of the binary and continuous parts of each submodel. Assuming linear change in each

outcome and random effects for the intercepts and time effects, for example, let $\mathbf{c}_{1i} = (a_{10i}, a_{11i}, b_{10i}, b_{11i})'$ for individual i be a vector of the random effects relating to the binary and continuous submodels, respectively, for the first measure, and similarly, $\mathbf{c}_{2i} = (a_{20i}, a_{21i}, b_{20i}, b_{21i})'$ be the comparable vector for the second measure. The random effects are assumed to be normally distributed with means equal to $\mathbf{0}$ and covariance matrix $\mathbf{\Phi}$:

$$\begin{pmatrix} \mathbf{c}_{1i} \\ \mathbf{c}_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \mathbf{\Phi} = \begin{pmatrix} \mathbf{\Phi}_1 & \\ \mathbf{\Phi}_{21} & \mathbf{\Phi}_2 \end{pmatrix}\right),$$

where $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$ are the covariance matrices of the latent change characteristics for the first and second variables, respectively, and $\mathbf{\Phi}_{21}$ is a nonsymmetric matrix of the covariances between the latent change characteristics of the different measures. Assuming unique variances and covariances among the random effects, for example, the distribution of the random effects is

$$\begin{pmatrix} a_{10i} \\ a_{11i} \\ b_{10i} \\ b_{11i} \\ a_{20i} \\ a_{21i} \\ b_{20i} \\ b_{21i} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \mathbf{\Phi} = \begin{bmatrix} \phi_{a100} & & & & & & & \\ \phi_{a11a10} & \phi_{a111} & & & & & & \\ \phi_{b10a10} & \phi_{b10a11} & \phi_{b100} & & & & & \\ \phi_{b11a10} & \phi_{b11a11} & \phi_{b11b10} & \phi_{b111} & & & & \\ \phi_{a20a10} & \phi_{a20a11} & \phi_{a20b10} & \phi_{a20b11} & \phi_{a200} & & & \\ \phi_{a21a10} & \phi_{a21a11} & \phi_{a21b10} & \phi_{a21b11} & \phi_{a21a20} & \phi_{a211} & & \\ \phi_{b20a10} & \phi_{b20a11} & \phi_{b20b10} & \phi_{b20b11} & \phi_{b20a20} & \phi_{b20a21} & \phi_{b200} & \\ \phi_{b21a10} & \phi_{b21a11} & \phi_{b21b10} & \phi_{b21b11} & \phi_{b21a20} & \phi_{b21a21} & \phi_{b21b00} & \phi_{b211} \end{bmatrix}\right)$$

Estimators and Algorithms

Once a statistical model is defined, an appropriate estimator and computational algorithm are chosen for specifying the rules for obtaining parameter estimates using certain computational procedures, based on observed data. In M*plus,* available maximum likelihood (ML) estimators vary according to the analytic model and outcome variable type. A two-part latent growth model for a single semicontinuous response variable requires a joint distribution of a binary and a continuous distribution, random effects at the second level of the model, and allowance for missing data. Three estimators, ML, MLR and MLF, are available in M*plus* to accommodate these features.

ML provides estimates with conventional standard errors. MLR is ML with standard errors that are robust to violations of the normality assumption. As described earlier, MLF is ML with standard errors that are calculated using first-order derivatives. These three estimators differ in their approach to approximating the Fisher information matrix. ML and MLR rely on an approximation to the Fisher information matrix that is based on second-order derivatives, whereas the approximation under MLF relies on first-order derivatives (Muthén, 1998–2004). For large samples, ML, MLR, and MLF produce equivalent results. These three estimators are available for a two-part latent growth model for one semicontinuous outcome, as specified earlier in this article.

In estimating a complex model such as a multivariate two-part latent growth model using M*plus,* however, ML and MLR are not available, and the MLF estimator is automatically implemented. Generally, MLF is recommended for computationally demanding estimation problems such as the current model, although research is needed to help understand the optimal number of integration points and increments for different kinds of problems.

Taking a likelihood-based approach, a goal in fitting a model to data is to obtain a set of parameter values that make the data most likely to be observed relative to any other parameter values. If latent variables are treated as random and parameters are fixed and unknown, inference is usually based on a marginal likelihood, the likelihood of the data given the latent variables, integrated over the latent variable distribution. Computationally, the parameters would maximize the likelihood function over its sample space.

M*plus* is a flexible statistical software package that can be used to fit such complex models using numerical integration for ML estimation. Researchers can choose the one that is considered most appropriate given the complexity of a model and computational resources. The default integration type in M*plus* is STANDARD, a rectangular (trapezoid) numerical integration method that uses 15 integration points per dimension. The integral of a function over a single dimensional space, the interval (a, b) for example, can be approximated by dividing the area bounded between a and b into multiple "rectangles" and summing the areas of the rectangles. This method can be extended to handle higher dimensions.

Given the computational demands that are inherent to the multivariate two-part latent growth model, a rectangular numerical integration method might not be feasible given the high-dimension computational problem; instead, MC integration (see Skrondal & Rabe-Hesketh, 2004) is applied. MC integration is an MC simulation-based approach that uses a random number generator to compute integrals, an approach that is considered preferable in dealing with estimation problems that require high-dimensional integration. Supposing we wish to compute an integral function $f(x)$ over sample space d, we could alternatively uniformly generate random samples over a space D, where d is a subset of D. The area of d is then estimated as the fraction

multiplied by the area of *D*. The larger the number of sample points, the more accurate the estimates. The main advantage of MC integration lies in its simplicity. The number of generated random samples, or integration points, can be specified by the researcher. For a suitably large number of integration points, this approach is considered to be very accurate. Users can implement MC integration in M*plus* by selecting the "ALGORITHM = MONTECARLO" option.

In M*plus,* a default of 500 random integration points is implemented for the MC integration algorithm. In some cases, one might wish to consider a series of analyses based on a range of integration points to take into account the sampling uncertainty of random integration points. A strategy for determining a suitable number of integration points is to plot parameter estimates across different numbers of integration points and select the number that demonstrates stability in the parameter estimates.

A SAS Macro Using MC Integration

To vary the number of random integration points when implementing the MC integration method using M*plus,* we developed an SAS macro *%M2Pfiles* to invoke M*plus* within SAS. The macro requires that users supply a data file and an M*plus* input script file that specifies the multivariate two-part latent growth model. The macro uses the SAS X command to repeatedly invoke M*plus* for the estimation of a model with a range of user-specified number of integration points. Output files are extracted corresponding to different integration points. This SAS macro is provided in Appendix A.

Although the notation used here for a two-part latent growth model and a multivariate two-part latent growth model is presented in long format (by the stacking of variables, as described earlier), the input data used in M*plus* is actually in wide format such that each individual has multiple variables rather than multiple data records. Examples of M*plus* input data and syntax codes are provided in Appendix B.

EXAMPLE: LONGITUDINAL MEASURES OF TIME-USE BEHAVIORS

Longitudinal measures of time-use behaviors were collected from a sample of children who participated in the Child Development Supplement (CDS), a part of the Panel Study of Income Dynamics (PSID) project. The PSID began in 1968 with a representative sample of U.S. children and families; details of the study can be found in Hill (1991). More recently, data on the PSID children and their families have been obtained through the CDS (Hofferth, Davis-Kean, Davis, & Finkelstein, 1997; Mainieri, 2006). Data were collected on average at 5-year intervals in 1997, 2002, and 2007, with response rates for the families within waves of 82%, 88%, and 76% for the time diaries, respectively.

For the PSID project, parents maintained a daily diary of their children's activities. The majority of families (70%) included in this subsample provided data for two children, and the remaining families included data for one child. Only data from the older of two siblings interviewed were included in this study. For the 5 twins included in the subsample, data for one of the two children were arbitrarily selected. Data for children who had reached the age of 18 years or older by the third wave were excluded from continued participation in the project. Using these selection criteria, a subset of 1,041 children were included for study here, of which 156 had measures for 1997 only, 76 for 2002 only, 15 for 2007 only, 573 for any 2 years, and 291 for all 3 years. The missing data were assumed to MAR, such that whether an individual was missing data could be related to the observed measures but not the missing data. Here, measures of time spent reading (Reading) and using a computer (Using a Computer), measured in hours and summed over 1 weekday and 1 weekend day for each study wave, were studied.

A Two-Part Latent Growth Model for Individual Time-Use Measures

A two-part latent growth model was fitted separately to Reading and Using a Computer measures to assess the form of growth in the probability of use and extent of conditional use in each behavior. This is a useful first step to obtain reasonable starting values for the larger multivariate model. Three forms of growth (i.e., no growth, linear growth, and quadratic growth) were fitted to each variable to assess the form of change as a function of the child's age. Age was centered at 13 years. In each model, the intercepts for each part of the growth model (i.e., that for the binary response and that for the continuous response) were specified to be random at the child level. The linear effects of time, but not the quadratic effect, were also assumed to be random. Given that children had at most three measures for a given variable, it is not possible to include more than two random effects in a model while assuming that the random effects covary. Estimates based on MLR and MLF were obtained using a rectangular and an MC integration procedure for comparison. Integration points ranging from 400 to 2,000 with increments of 200 points were selected for use with the MC integration procedure. Given that only MLF was available for fitting the large, more complex model, estimates of these simpler models produced by MLR and MLF were compared in this initial step in the data analysis to document any important differences, particularly with regard to the estimated standard errors.

A Multivariate Two-Part Latent Growth Model for Time-Use Data

Our goal in jointly modeling the time-use outcomes was to investigate (a) the probability, (b) the extent, and (c) the relationship between the probability and extent of reading and using a computer at home. The Reading and Using a Computer measures collected from the same individual might share some common characteristics. Thus, these outcomes are likely to be correlated. A model was formulated based on the results from the previous analyses in which each time-use variable was considered individually. Estimates based on MLF were obtained through the MC integration method implemented in M*plus,* with integration points ranging from 400 to 2,000 in 200-point increments. Statistical tests were performed at the .05 level.

RESULTS

Descriptive statistics of age in years at each year of measurement are given in Table 1. The proportion of zeros and descriptive statistics for the positive values of Reading and Using a Computer are also provided. Figure 1 displays the histograms of Reading and Using a Computer over the three measurement occasions.

Single-Outcome Two-Part Latent Growth Model

Indexes of model fit (i.e., the Akaike Information Criterion [AIC] and the Bayesian Information Criterion [BIC]; values not reported) indicated that change in each aspect of Reading and Using a Computer was best described by a linear growth model with random intercepts in each submodel of the binary and conditional continuous outcome. The random intercepts of the two model parts were allowed to covary. As shown in Tables 2 and 3 for Reading and Using a Computer, respectively, estimates of the fixed effects, variances of the Level 1 errors, and variances of the random intercepts at Level 2 produced by rectangular and MC integration methods using MLF were in close agreement. The *SE*s using different estimators differed only slightly (*SE*s from MLR are available on request).

A test of the assumption of homogeneity of variance for the occasion-specific errors for Reading and Using a Computer indicated differences in the variances across time, $\chi^2(2) = 74.46, p < .01$ for Reading; $\chi^2(2) = 72.86, p < .01$ for Using a Computer. Thus, a model that allowed for unique variances at each time point was retained.

Multivariate Two-Part Latent Growth Model

A multivariate two-part latent growth model was fitted to the two time-use variables studied. This model could not be estimated with rectangular integration using an Intel dual-core processor 2.8 GHz, 4 GB RAM computer on a 32-bit operating system, as the computational capacity was exceeded. The final results were based on MLF through MC integration. A path diagram of the fitted model is presented in Figure 2. A plot of one of the parameters is presented in Figure 3. In this plot, the values of the parameters (vertical axis) are plotted against the number of MC integration points (horizontal axis). As can be seen, the parameter estimate does not vary appreciably across the number of integration points. Parameter estimates and standard errors based on MLF using 1,600 integration points are provided in Table 4. Parameter estimates and standard errors do not differ appreciably from the corresponding values obtained from the analyses in which measures of Reading and Using a Computer were considered individually.

For time spent reading, the estimated log odds that a child will read at age 13 is $\alpha_{10} = -1.12(SE = 0.09)$,

TABLE 1
Descriptive Statistics for a Subsample of Child Development Supplement Children

	n	*% of Zeros*	*n*	*M*	*SD*	*Skewness*
Age						
1997			900	7.59	1.46	−0.10
2002			827	13.11	1.50	−0.10
2007			399	16.31	0.88	−0.09
Time spent Reading						
	Proportion of Reading			Conditional extent of Reading		
1997	900	56.56%	391	1.03	0.91	1.82
2002	827	68.56%	260	1.45	1.44	2.59
2007	399	78.7%	85	1.21	1.10	1.80
Time spent Using a Computer						
	Proportion of Using a Computer			Conditional Extent of Using a Computer		
1997	900	87.67%	111	1.68	1.29	1.73
2002	827	62.39%	311	2.39	2.18	1.80
2007	399	40.35%	238	3.14	2.83	2.36

Note. $N = 1{,}041$.

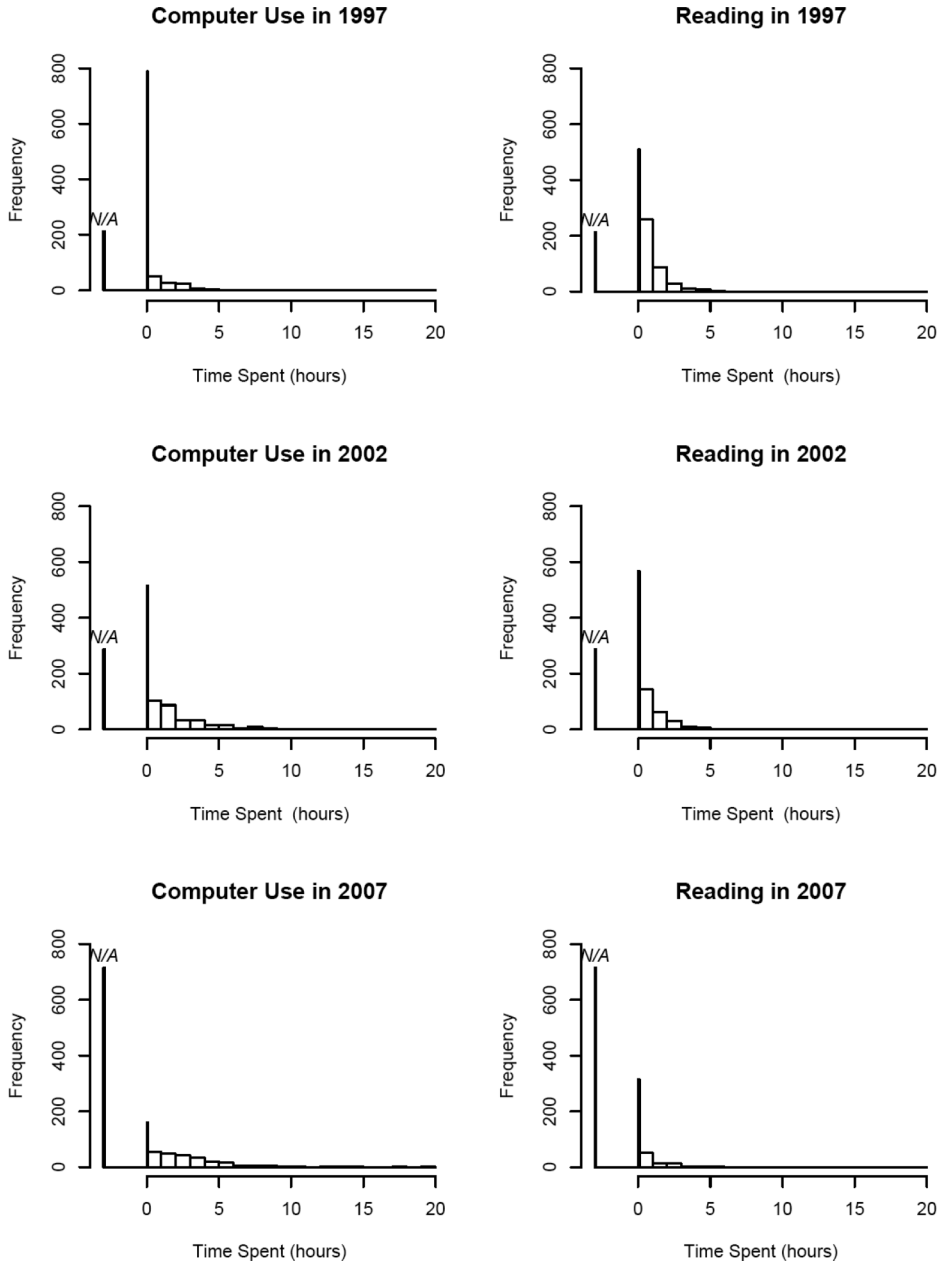


FIGURE 1 Histograms of time-use variables over repeated measures. Time spent on Using a Computer and Reading is calculated as a sum over 1 weekday and 1 weekend day. The missing values are denoted as N/A.

corresponding to a probability of $(1/(1 + \exp(1.12)) = 0.25$. The log odds that a child will read declined linearly with age ($\alpha_{11} = -0.15, SE = 0.02$). Conditional that a child did read, the estimated mean time spent reading at age 13 was $\beta_{10} = 1.16$, suggesting an average of slightly more than 1 hour spent reading across the 1 weekday and 1 weekend day. Additionally, the estimated time spent reading increased slightly ($\beta_{11} = 0.04, SE = 0.02$) on average with age.

Next, the estimated probability that a child used a computer at age 13 was $\alpha_{20} = -0.65$ ($SE = 0.07$), corresponding to a probability of $(1/(1 + \exp(0.65)) = 0.34$. On average, the log odds that a child used a computer increased with age ($\alpha_{21} = 0.28$, $SE = 0.02$). Conditional that a child did use a computer, the estimated mean time spent using a computer at age 13 was $\beta_{20} = 2.41$ ($SE = 0.18$), an estimate of nearly 2.5 hours across the 2 days of measurement. On average, time spent using a computer increased at a rate of about $\beta_{21} = 0.16$ ($SE = 0.03$) with age.

At the occasion level, previous analyses suggested differences between the variances of the time-specific errors of both Reading and Using a Computer across the three

TABLE 2
Parameter Estimates for a Two-Part Latent Growth Model for Repeated Measures of Time Spent on Reading

	Rectangular Integration		*Monte Carlo Integration*	
Fixed Effect	*Estimate*	*SE*	*Estimate*	*SE*
Logit submodel				
Intercept α_0	−1.11	0.09	−1.01	0.09
Age α_1	−0.15	0.03	−0.15	0.02
Continuous submodel				
Intercept β_0	1.16	0.11	1.14	0.11
Age β_1	0.04	0.02	0.03	0.02
Variances of Level 1 residuals and Level 2 random intercept				
Level 1				
σ^2_{97}	0.61	0.09	0.61	0.09
σ^2_{02}	1.89	0.14	1.89	0.14
σ^2_{07}	1.11	0.16	1.10	0.15
Level 2				
ϕ_{00}	1.43	0.30	1.38	0.30
ϕ_{10}	0.20	0.12	0.21	0.11
ϕ_{11}	0.21	0.08	0.21	0.08

Note. Results from Monte Carlo integration are based on 1,000 integration points with maximum likelihood estimates with standard errors that are approximated by first-order derivatives. ϕ_{00} = variance of the random intercept of the binary model part; ϕ_{10} = covariance between the random intercept of the binary and continuous model part; ϕ_{11} = variance of random intercept of the continuous model part.

TABLE 3
Parameter Estimates for a Two-Part Latent Growth Model for Repeated Measures of Time Spent on Using a Computer

	Rectangular Integration		*Monte Carlo Integration*	
Fixed Effect	*Estimate*	*SE*	*Estimate*	*SE*
Logit submodel				
Intercept α_0	−0.66	0.07	−0.66	0.07
Age α_1	0.29	0.02	0.29	0.02
Continuous submodel				
Intercept β_0	2.39	0.18	2.40	0.18
Age β_1	0.16	0.03	0.16	0.03
Variances of Level 1 residuals and Level 2 random intercept				
Level 1				
σ^2_{97}	1.36	0.35	1.36	0.35
σ^2_{02}	4.14	0.43	4.15	0.43
σ^2_{07}	7.64	0.50	7.65	0.50
Level 2				
ϕ_{00}	0.92	0.29	0.91	0.29
ϕ_{10}	0.12	0.24	0.09	0.23
ϕ_{11}	0.42	0.34	0.41	0.34

Note. Results from Monte Carlo integration are based on 1,000 integration points with maximum likelihood estimates with standard errors that are approximated by first-order derivatives. ϕ_{00} = variance between the random intercept of the binary and continuous model part; ϕ_{10} = covariance between the random intercept of the binary and continuous model part; ϕ_{11} = variance of random intercept of the continuous model part.

measurement occasions. Also within occasion, a likelihood ratio test for a possible covariation between Reading and Using a Computer suggested that the two behaviors did not covary within occasion, assuming the covariance was constant across time, $\chi^2(1) = 1.30$, $p = .25$.

At the individual level, the standard error of the estimated variance of the random intercept of the continuous submodel for Using a Computer was not large relative to the size of the estimated variance ($\Phi_{44} = 0.37, SE = 0.34$), suggesting that a model with a random intercept might not be appreciable. We conducted a test of the null hypothesis that the variance of the random intercept of the continuous submodel was zero, which is equivalent to testing the null hypothesis that four parameters, namely the variance of the random intercept and its covariances with the remaining three random effects, are all zero. It is important to note that this null hypothesis places the variance parameter on the boundary of the parameter space, so that the p value for the standard chi-square approximation for the likelihood ratio test tends to be conservative (Pinheiro & Bates, 2000). The appropriate null distribution in this case is a mixture of $\chi^2(4)$ and $\chi^2(3)$, with each having an equal weight of 0.5 (see Verbeke & Molenberghs, 2000). Given the log-likelihood of the null model is –4992.67, and of the alternative model is –4990.09, twice the difference in these likelihoods is 5.16. Adopting this rule, the p value for the comparison of two nested models can be calculated as

$$p = \frac{1}{2}p(\chi^2(4) \geq 5.16) + \frac{1}{2}p(\chi^2(3) \geq 5.16) = 0.62.$$

From this we concluded that the variation in time spent on Using a Computer at age 13 was not different from 0.

An estimated covariance of 0.72 ($SE = 0.14$), $\chi^2(1) = 29.90$, $p < .001$, between the random intercepts relating to the log odds of Reading and the log odds of Using a Computer was statistically different from zero, suggesting that at 13 years of age, the likelihood that a child will read is positively related to the likelihood that a child will use a computer. No other association was detected between the latent growth characteristics of Reading and Using a Computer.

DISCUSSION

Semicontinuous variables are common in many areas of behavioral research. Such variables are characterized by a response scale in which some scores take on a single value (often zero) and the remaining scores are continuous (usually positive). Given this response distribution, it might be useful to distinguish between the two features of the response in studying whether or not individuals engage

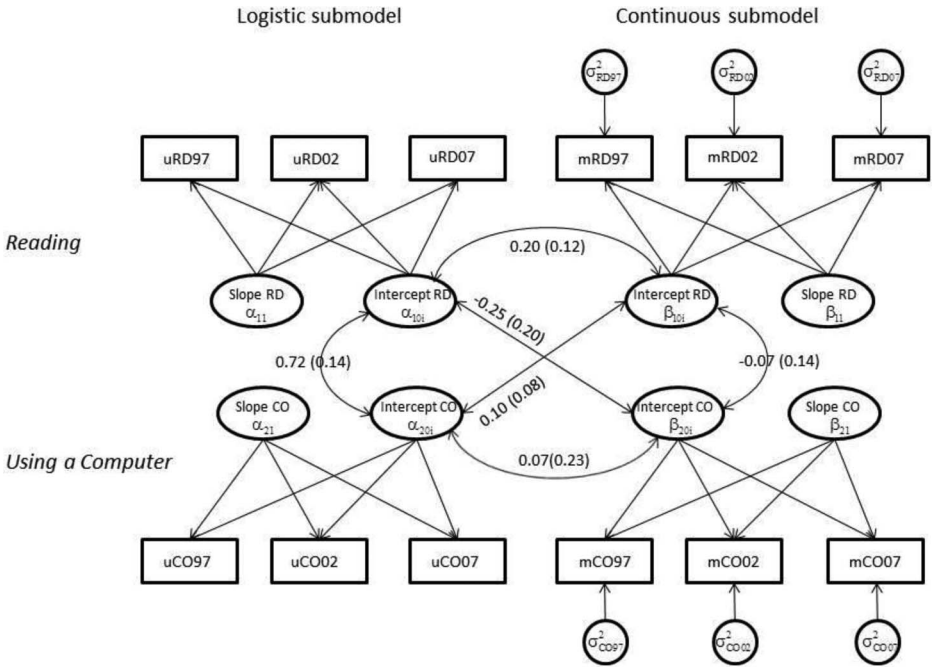


FIGURE 2 Path diagram of a multivariate two-part latent growth model for longitudinal measures of Reading and Using a Computer. uRD97, uRD02, uRD07 are binary measures of Reading in 1997, 2002, and 2007, respectively; mRD97, mRD02, mRD07 are nonzero measures of Reading in 1997, 2002, and 2007, respectively; similar measures are provided for measures of Using a Computer in 1997, 2002 and 2007, respectively; Intercept RD α_{10i} = random intercept of Reading binary model part; Slope RD α_{11} = slope of Reading binary mode part; Intercept RD β_{10i} = random intercept of Reading continuous model part; Slope RD β_{11} = slope of Reading continuous model part; Intercept CO α_{20i} = intercept of Using a Computer binary model part; Slope CO α_{21} = slope of Using a Computer binary model part; Intercept CO β_{20i} = intercept of Using a Computer continuous model part; Slope CO β_{21} = slope of Using a Computer continuous model part; σ^2_{RD97}, σ^2_{RD02}, and σ^2_{RD07} are variance of residuals of Reading continuous model part in 1997, 2002, and 2007, respectively; similar measures are provided for measures of Using a Computer in 1997, 2002 and 2007, respectively.

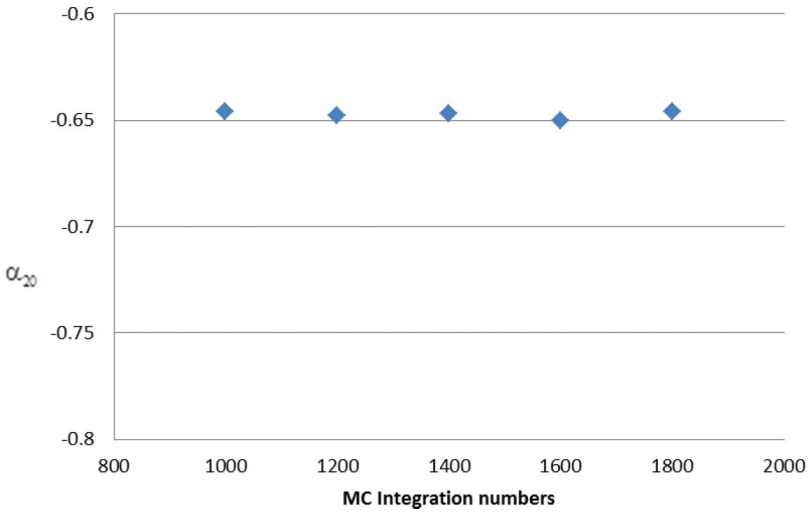


FIGURE 3 Stable parameter estimates across a multivariate two-part latent growth model across different integration points. α_{20} = intercept of Using a Computer binary model part; MC = Monte Carlo. (Figure appears in color online.)

in a particular behavior (e.g., adolescent alcohol use), and if they do, to measure the extent to which they engage (e.g., amount of alcohol use). A two-part latent growth model might be considered for a single longitudinal semicontinuous variable.

This article develops a multivariate two-part latent growth model for the study of two variables that individually follow a two-part latent growth model. Thus, the model considers the joint distributions of two binary and two continuous response variables. The statistical software package M*plus*

TABLE 4
Parameter Estimates for a Multivariate Two-Part Latent Growth Model for Repeated Measures of Time Spent on Reading and Using a Computer

	Reading			*Using a Computer*	
	Estimate	*SE*		*Estimate*	*SE*
Fixed effect					
Logistic submodel					
Intercept α_{10}	−1.12	0.09	Intercept α_{20}	−0.65	0.07
Age α_{11}	−0.15	0.02	Age α_{21}	0.28	0.02
Continuous Submodel					
Intercept β_{10}	1.16	0.10	Intercept β_{20}	2.41	0.18
Age β_{11}	0.04	0.02	Age β_{21}	0.16	0.03
Level 1 residual variance					
σ^2_{RD97}	0.60 (0.09)		σ^2_{CO97}	1.36	0.35
σ^2_{RD02}	1.89 (0.14)		σ^2_{CO02}	4.16	0.43
σ^2_{RD07}	1.11 (0.16)		σ^2_{CO07}	7.83	0.50
Level 2 covariance matrix					

$$\Phi = \begin{bmatrix} 1.43\,(0.30) & & & \\ 0.20\,(0.12) & 0.21\,(0.08) & & \\ 0.72\,(0.14) & 0.10\,(0.08) & 0.78\,(0.26) & \\ -0.25\,(0.20) & -0.07\,(0.14) & 0.07(0.23) & 0.37\,(0.34) \end{bmatrix}$$

Note. Results from Monte Carlo integration with maximum likelihood estimates with standard errors that are approximated by first-order derivatives are based on 1,600 integration points. The Level 2 covariance matrix is a symmetric matrix; on the diagonal are the variance of random intercepts of the binary model part of Reading, continuous model part of Reading, binary model part of Using a Computer, and continuous model part of Using a Computer, respectively. The off-diagonal components are the covariances between each of the four random effects.

can be used to estimate such a model using MC integration. This article also develops an SAS macro to invoke M*plus* to fit a series of multivariate two-part latent growth models using a range of numbers of integration points for MC. Final results can be obtained by assessing the stability of parameter estimates across different numbers of integration points. Another approach is the quasi-Newton optimization of the likelihood approximated by adaptive Gaussian quadrature that can be implemented using SAS PROC NLMIXED (an SAS example code is provided by Tooze et al., 2002). The multidimensional integration that is necessary to evaluate a likelihood function is difficult, however, in fitting a two-part latent growth curve model, with greater difficulty in a multivariate case. This could present a great computational challenge if using PROC NLMIXED to fit such a model.

In the illustrative example, two estimation methods, MLR and MLF, were used to fit a univariate two-part latent growth model for measures of time spent reading and using a computer, respectively. Both estimators provide similar estimates with slight differences in the estimated standard errors. In small samples, ML estimates with robust standard errors are often preferred to those obtained using MLF given the robustness of MLR to the normality assumption. In some cases (e.g., a multivariate two-part latent growth model), however, ML with robust standard errors cannot be computed given that a high-dimensional integral is required to compute a second-order derivative. Thus, MLF might be the only option.

Given the complexity of a multivariate two-part growth curve model, MC integration was applied in estimating the correlation between two longitudinal semicontinuous variables in a study of time use in children. For MC integration, integration points from 400 to 2,000 were selected; however, the models failed to converge for integration points of 400, 600, 800, and 2,000. We also selected the M*plus* default value of 500 and the model failed to converge. Finally, the results from MC integration based on 1,600 integration points are presented in Table 4 for this study. A plot can be considered to evaluate the stability of a parameter estimate under different integration numbers. An unstable estimate of a parameter is represented by a fluctuant band, as shown in Figure 4. A plot of unstable estimates could present an upward or downward trend. In this study, all of the parameter estimates and model fit indices achieved stability, with the exception of the covariance between the random intercepts of the binary submodel and the continuous submodel of Using a Computer. We also found that, consistently across different integration points, the estimated standard errors were large relative to the size of the corresponding covariance. Given that the remainder of the parameter estimates was stable and the interpretation of this parameter will not change across different integration points, we decided to report the final results based on the 1,600 integration points. Additionally, one fundamental disadvantage of MC integration is that its accuracy increases only as the square root of the number of random integration points increases (Teukolsky, Vetterling, & Flannery, 2007). Therefore, if the accuracy requirements are modest, or if the computation budget is large, then the technique is highly recommended as one of great generality. More studies need to be done to explore the range of optimal numbers of integration points and increments.

Finally, empirical studies in the social sciences often encounter data in which respondents born in different years are observed at multiple points in time (e.g., Glenn, 2007). This often presents a challenge of separating age, cohort, and time effects because two or three of these effects are always confounded. We simplified our demonstration by studying age effects in a multivariate two-part latent growth model, although the effects of age and time were confounded. Further, we assumed that the cohort effects in the growth trajectories were negligible. Given differences in baseline ages for children in this study, measuring change as a function of age required the assumption of age convergence, which implies that younger people and older people differ only by age, and that between-person, cross-sectional age effects are equivalent to within-person, longitudinal aging effects (Hoffman, Hofer, & Sliwinski, 2011).

Miyazaki and Raudenbush (2000) proposed an approach for studying multiple cohorts in a longitudinal design to address an interaction between participant's age and cohort.

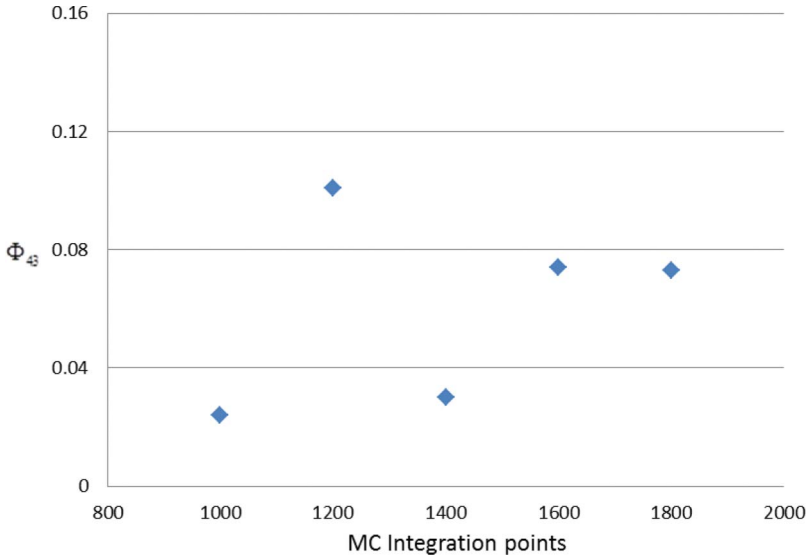


FIGURE 4 Unstable parameter estimates of a multivariate two-part latent growth model across different integration points. Φ_{43} = covariance between random intercepts of Using a Computer. MC = Monte Carlo. (Figure appears in color online.)

Using this approach, the coefficients that describe growth at the individual level can be specified as a function of cohort membership, allowing, for instance, for the slope of the growth trajectory to vary on average between cohorts. In this study, for instance, cohorts could be formed based on year of birth. Selecting one birth cohort as the comparison or reference group, indicator variables would be created to represent the different birth cohorts. These indicator variables would then enter the model at the second level to test whether there are cohort differences in the characteristics used to describe change.

ACKNOWLEDGMENT

This work was supported in part by the William T. Grant Foundation (200601036).

REFERENCES

Altheide, D. L. (1997). Media participation in everyday life. *Leisure Sciences*, *19*, 17–29.

Blozis, S. A. (2004). Structured latent curve models for the study of change in multivariate repeated measures. *Psychological Methods*, *9*, 334–353.

Blozis, S. A. (2007). A second-order structured latent curve model for longitudinal data. In K. van Montfort, J. Oud, & A. Satorra (Eds.), *Longitudinal models in the behavioral and related sciences* (pp. 189–214). Mahwah, NJ: Erlbaum.

Blozis, S. A., Feldman, B., & Conger, R. (2007). Adolescent alcohol use and adult alcohol disorders: A two-part random-effects model with diagnostic outcomes. *Drug and Alcohol Dependence*, *88*, 85–96.

Brown, E. C., Catalano, R. F., Fleming, C. B., Haggerty, K. P., & Abbott, R. D. (2005). Adolescent substance use outcomes in the Raising Healthy Children Project: A two-part latent growth curve analysis. *Journal of Consulting and Clinical Psychology*, *73*, 699–710.

Duncan, S., Duncan, T. E., & Strycker, L. A. (2006). Alcohol use from ages 9–16: A cohort-sequential latent growth model. *Drugs and Alcohol Dependence*, *81*, 71–81.

Glenn, N. D. (2007). Age, period, and cohort effects. In G. Ritzer (Eds.), *Blackwell encyclopedia of sociology* (pp. 52–56). Malden, MA: Blackwell.

Hall, D. B. (2000). Zero-inflated Poisson and binominal regression with random effects: A case study. *Biometrics*, *56*, 1030–1039.

Hill, M. S. (1991). *The panel study of income dynamics: A user's guide*. Thousand Oaks, CA: Sage.

Hofferth, S., Davis-Kean, P. E., Davis, J., & Finkelstein, J. (1997). *The Child Development Supplement to The Panel Study of Income Dynamics: 1997 User Guide*. Retrieved from http://psidonline.isr.umich.edu/CDS/usergd.html

Hoffman, L., Hofer, S. M., & Sliwinski, M. J. (2011). On the confounds among retest gains and age-cohort differences in the estimation of within-person change in longitudinal studies: A simulation study. *Psychology and Aging*, *26*, 778–791.

Hoffmann, J. P., Cerbone, F. G., & Su, S. S. (2000). A growth curve analysis of stress and adolescent drug use. *Substance Use and Misuse*, *35*, 687–716.

Liang, N. M., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*, 13–22.

Liu, L., Ma, J. Z., & Johnson, B. A. (2008). A multi-level two-part random effects model, with application to an alcohol-dependence study. *Statistics in Medicine*, *27*, 3528–3539.

MacCallum, R. C., Kim, C., Malarkey, W. B., & Kiecolt-Glaser, J. K. (1997). Studying multivariate change using multilevel models and latent curve models. *Multivariate Behavioral Research*, *32*, 215–253.

Mainieri, T. (2006). *The panel study of income dynamics child development supplement: User guide for CDS II*. Available from http://psidonline.isr.umich.edu/CDS/cdsii_userGd.pdf

Meredith, W., & Tisak, J. (1984, June). *Tuckerizing curves*. Paper presented at the annual meeting of the Psychometric Society, Santa Barbara, CA.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, *55*, 107–122.

Miyazaki, Y., & Raudenbush, S. W. (2000). Test for linkage of multiple cohorts in an accelerated longitudinal design. *Psychological Methods*, *5*, 44–63.

Muthén, B. O. (2001). *Two-part growth mixture modeling*. Unpublished manuscript.

Muthén, B. O. (1998–2004). *M*plus *technical appendices*. Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (1998–2010). *M*plus *user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Olsen, M. K., & Schafer, J. L. (2001). A two-part random effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association*, *96*, 730–745.

Petras, H., Nieuwbeerta, P., & Piquero, A. R. (2010). Participation and frequency during criminal careers over the life span. *Criminology*, *48*, 607–637.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York, NY: Springer.

Reinsel, G. (1992). Multivariate repeated-measurement or growth curve models with multivariate random-effects covariance structure. *Journal of the American Statistical Association*, *77*, 190–195.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. New York, NY: Chapman & Hall/CRC.

Su, L., Tom, B. D. M., & Farewell, V. T. (2009). Bias in 2-part mixed models for longitudinal semicontinuous data. *Biostatistics*, *10*, 374–389.

Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2007). Random numbers. In *Numerical recipes: The art of scientific computing* (pp. 340–418). Hong Kong: William H. Press.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, *26*, 24–36.

Tooze, J. A., Grunwald, G. K., & Jones, R. H. (2002). Analysis of repeated measures data with clumping at zero. *Statistical Methods in Medical Research*, *11*, 341–355.

Vazsonyi, A. T., & Keiley, M. K. (2007). Normative developmental trajectories of aggressive behaviors in African American, American Indian, Asian American, Caucasian, and Hispanic children and early adolescents. *Journal of Abnormal Child Psychology*, *35*, 1047–1062.

Verbeke, G., & Davidian, M. (2008). Joint models for longitudinal data: Introduction and review. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis: Handbooks of modern statistical methods* (pp. 319–326). Boca Raton, FL: Chapman & Hall/CRC Press.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York, NY: Springer Verlag.

Wang, L., Zhang, Z., McArdle, J. J., & Salthouse, T. A. (2008). Investigating ceiling effects in longitudinal data analysis. *Multivariate Behavioral Research*, *43*, 476–496.

Weaver, S. R., Cheong, J., MacKinnon, D., & Pentz, M. A. (2011). Investigating ethnic differences in adolescent alcohol use and peer norms using semi-continuous latent growth models. *Alcohol and Alcoholism*, *46*(4), 1–7.

Witkiewitz, K., & Masyn, K. (2008). Drinking trajectories following an initial lapse. *Psychology of Addictive Behaviors*, *22*, 157–167.

APPENDIX A
SAS MACRO FOR IMPLEMENTING A MULTIVARIATE TWO-PART LATENT GROWTH MODEL USING MC INTEGRATION

```
* This document contains the SAS syntax to remotely access Mplus to fit a multivariate two-part latent
growth model;
* The Mplus syntax to fit a multivariate two-part latent growth model is in Mplus version 4 format, in
which a new binary and continuous variable is pre-created by the user for each semi-continuous variable.
Since Mplus version 5, a new statement "Data Twopart" is introduced to define a semi-continuous variable.
A binary and continuous variable will be automatically created, and a logarithmic transformation will
be applied to the continuous variable by default. To be consistent to the mathematical notation in the
article, an example code in Mplus version 4 format is provided in Appendix A;
** Model: Multivariate TWO-PART LATENT GROWTH MODEL;
* Specify the path where the estimates will be saved;
%let path = E:\;
libname new "&path";
%macro M2Pfiles(minnumb = , maxnumb = , bynumb = ) ;
%do intnumb = &minnumb %to &maxnumb %by &bynumb;
* Step 1: save each replication as a separate temp .txt file with the same file name;
* Specifying the macro variable &intnumb allows the user to provide a range of integration numbers using
MC integration;
data mplus;
file "d:\simcode.txt ";
* Type in Mplus code. Attention, only code written in lower case will work;
put@1 "title: joint distribution of computer use and rd, ";
put@1" binary + continuous data , 2 level; ";
put@1" data: file is empiricalstudy.dat ; ";
put@1" format is free; ";
put@1 "variable: ";
put@1 " names are id68pn male agec97 agec02 agec07 ";
put@1 " uco97 uco02 uco07 mco97 mco02 mco07 ";
put@1 " urd97 urd02 urd07 mrd97 mrd02 mrd07 agepm1 agepm2 ; ";
put@1 "missing is . ; ";
put@1 " usevariables are agec97 agec02 agec07 mco97 mco02 mco07 ";
put@1 " uco97 uco02 uco07 urd97 urd02 urd07 mrd97 mrd02 mrd07; ";
put@1 " categorical are uco97 uco02 uco07 urd97 urd02 urd07 ; ";
put@1 "tscores = agec97 agec02 agec07 ; ";
put@1 "analysis: type = missing random; ";
put@1 " integration = montecarlo(&intnumb); ";
put@1 " estimator = mlf; ";
put@1 "model: ";
put@1 " !! co data; ";
put@1 " ! binary part; ";
put@1 " iuco suco | uco97 uco02 uco07 at agec97 agec02 agec07 ; ";
put@1 " suco@0; ";
put@1 " ! continuous part; ";
put@1 " imco smco | mco97 mco02 mco07 at agec97 agec02 agec07 ; ";
put@1 " smco@0; ";
put@1 " mco97 mco02 mco07 * 4.74 ; ";
put@1 " [suco * 0.29];[imco * 2.39]; [smco*0.16]; ";
put@1 " iuco * 0.92; imco * 0.42; ";
put@1 " !! rd data; ";
put@1 " ! binary part; ";
put@1 " iurd surd | urd97 urd02 urd07 at agec97 agec02 agec07 ; ";
put@1 " surd@0; ";
put@1 " ! continuous part; ";
put@1 " imrd smrd | mrd97 mrd02 mrd07 at agec97 agec02 agec07 ; ";
put@1 " smrd @0; ";
put@1 " mrd97 mrd02 mrd07 *1.149 ; ";
put@1 " [surd * -0.15]; [imrd * 1.14]; [smrd * 0.03]; ";
put@1 " iurd * 1.38; imrd * 0.21; ";
put@1 " ! the random intercepts of the two parts are correlated; ";
put@1 " imco with iuco * 0.78; ";
put@1 " iurd with iuco * 0.72; ";
put@1 " iurd with imco * -0.25; ";
```

```
put@1 " imrd with iuco * 0.10; ";
put@1 " imrd with imco * 0.06; ";
put@1 " imrd with iurd * 0.12; ";
put@1 " output: tech1 tech4; ";
;
run;
* Step 2: run one replication at a time in Mplus remotely from SAS;
* & inputname: indicate the file contains the input Mplus code;
* & outputname: indicate the file contains the output Mplus code;
%let inputname=d:\simcode.txt;
%let outputname=d:\simcode&intnumb..out;
option noxwait xsync;
data _null_;
x "mplus &inputname &outputname";
run;
quit;
*Step 3: extract the estimators from each replication;
%let filename=d:\simcode&intnumb..out;
filename ABC "&filename";
** Read the parameter estimates from simcode&intnumb..out;
data modelfit;
infile ABC length=lg;
input @;
input @1 eachline $varying200. lg;
x1 = index(eachline,"H0 Value");
x2 = index(eachline,"Number of Free Parameters ");
x3 = index(eachline,"Akaike (AIC)");
x4 = index(eachline,"Bayesian (BIC)");
x5 = index(eachline,"Sample-Size Adjusted BIC");
if x1=11 or x2 =11 or x3=11 or x4 = 11 or x5 = 11 ;
run;
data modelfit2;
set modelfit;
varname = compress(substr(eachline,1,39));
value = substr(eachline,39,20);
keep varname value;
run;
PROC TRANSPOSE data = modelfit2 out = modelfit3;
id varname;
var value;
run;
data modelfit4;
set modelfit3;
ind = &intnumb;
H0LikeValue = input(H0Value, 20.4);
NofParms = input(NumberofFreeParameters, 20.0);
AIC = input(Akaike_AIC_, 20.4);
BIC = input(Bayesian_BIC_,20.4);
AdjBIC = input(Sample_SizeAdjustedBIC, 20.4);
drop H0Value NumberofFreeParameters Akaike_AIC_ Bayesian_BIC_ Sample_SizeAdjustedBIC _name_;
run;
*****************************************************;
data estimates;
infile ABC length=lg;
input @;
input @1 eachline $varying200. lg;
d1=index(eachline,"IMCO");
d2=index(eachline,"IMRD");
d3=index(eachline,"IUCO");
d4=index(eachline,"IURD");
d5=index(eachline,"SMCO");
d6=index(eachline,"SMRD");
d7=index(eachline,"SUCO");
d8=index(eachline,"SURD");
d9=index(eachline,"MCO97");
d10=index(eachline,"MCO02");
```

```
d11=index(eachline,"MCO07");
d12=index(eachline,"MRD97");
d13=index(eachline,"MRD02");
d14=index(eachline,"MRD07");
d15=index(eachline,"UCO97$1");
d16=index(eachline,"URD97$1");
if d1 = 5 or d2 = 5 or d3 = 5 or d4 = 5 or d5 = 5 or d6 = 5 or d7 = 5 or d8 = 5 or d9 = 5 or d10 = 5
or d11 = 5 or d12 = 5 or d13 = 5 or d14 =5 or d15 = 5 or d16 = 5;
run;
data estiamtes2;
set estimates;
if _N_ = 15 then delete ;
if _N_ = 16 then delete ;
if _N_ = 17 then delete ;
if _N_ = 18 then delete ;
if _N_ = 19 then delete ;
if _N_ = 20 then delete ;
Estimates = substr(eachline,12,21) ;
SE = substr(eachline,30,12) ;
Walds = substr(eachline,40,12) ;
keep varname Estimates SE Walds ;
run;
**************************;
*So far, all the values are still Strings (characteristic type);
data estimates3;
set estiamtes2;
Estimates1 = input(Estimates, 22.4);
SE1 = input(SE , 22.4);
Walds1 = input(Walds, 22.4);
drop Estimates SE Walds Std StdXY;
RUN;
proc transpose data=estimates3 out = estimate4 LET ; RUN;
data estimate5;
set estimate4;
ind = &intnumb;
rename col1 = IUCO_IMCO;
rename col2 = IURD_IUCO;
rename col3 = IURD_IMCO;
rename col4 = IMRD_IUCO;
rename col5 = IMRD_IMCO;
rename col6 = IMRD_IURD;
rename col7 = m_IUCO;
rename col8 = m_SUCO;
rename col9 = m_IMCO;
rename col10 = m_SMCO;
rename col11 = m_IURD;
rename col12 = m_SURD;
rename col13 = m_IMRD;
rename col14 = m_SMRD;
rename col15 = thr_CO;
rename col16 = thr_RD;
rename col17 = var_IUCO;
rename col18 = var_SUCO;
rename col19 = var_IMCO;
rename col20 = var_SMCO;
rename col21 = var_IURD;
rename col22 = var_SURD;
rename col23 = var_IMRD;
rename col24 = var_SMRD;
rename col25 = res_CO97;
rename col26 = res_CO02;
rename col27 = res_CO07;
rename col28 = res_RD97;
rename col29 = res_RD02;
rename col30 = res_RD07;
run;
```

```
*Combine the model fit and estimators of the same replication;
data combine;
merge estimate5 modelfit4;
by ind;
run;
*Step 4: save the parameter estimates and model fit indices;
proc append base = new.mpluscombine_
COREfeml force;
run;
*end intnumb loop;
%end;
%Mend M2Pfiles ;
*run the macro with user-specified parameters;
%M2Pfiles(minnumb = 400, maxnumb = 2000 , bynumb = 200 ) ;
```


APPENDIX B
CREATING AN M*PLUS* INPUT DATA SET FOR A TWO-PART LATENT GROWTH MODEL

M*plus* data sets containing multilevel or hierarchical data can be organized in one of two ways: (a) multiple record data set (referred to as long format), in which data are structured in multiple rows per individual, and (b) multiple variable data set (referred to as wide format), in which one row with multiple variables is used to record measures from an individual on multiple occasions. To use M*plus* for a two-part latent growth model, a wide format data set is needed. One can use the "Data Twopart:" statement as initially introduced in M*plus* Version 5, which is shown in Example Code 1. For any version of M*plus,* the script can be written as though it is a case of structural equation modeling, as shown in Example Code 2. In a free format M*plus* input data set, a missing record is coded as a comma and each variable is separated by an empty column. Codes 1 and 2 have the same function. They both prepare the data.

Example Code 1 for M*Plus* Version 5 and Higher

There are nine variables for each observation. The "Names = " statement specifies the original semicontinuous variable. M*plus* will automatically recode the variable into a binary and a continuous outcome, the names of which are specified in the "Binary = " and the "Continuous = " statements, respectively. Note that if the "Data Twopart:" statement is specified, M*plus* applies a natural log transformation to the values of the continuous model part by default. Users can specify the "Transform = " statement to override the default. In the "Variable:" command, the "Names =" statement specifies the original variables in the input data set. The "Usevariables =" statement defines the variables used in the model specification. The binary or categorical dependent variables in model estimation are provided in the "Categorical = " statement. In our example, these nine variables from the input data set are age centered at 13 years old across three measures in columns 1 through 3, original semicontinuous measures of Using a Computer in columns 4 through 6, and original semicontinuous measures of Reading in columns 7 through 9.

```
Data Twopart:
    Names = CO97 CO02 CO07 RD97 RD02 RD07;
    Binary = uCO97 uCO02 uCO07 uRD97 uRD02 uRD07;
    Continuous = mCO97 mCO02 mCO07 mRD97 mRD02 mRD07;
    Transform = NONE;
Variable:
    Names = agec97 agec02 agec07 CO97 CO02 CO07 RD97 RD02 RD07;
    Usevariables = agec97 agec02 agec07 uCO97 uCO02 uCO07 mCO97 mCO02 mCO07
    uRD97 uRD02 uRD07 mRD97 mRD02 mRD07;
    Categorical = uCO97 uCO02 uCO07 uRD97 uRD02 uRD07;
Input data:
    −4.79    0.82    5.17    2.63    4.18    .       0.65    0       .
    −6.82    −1.11   3.17    0       0       4.42    0.17    0       0
    −6.79    −1.27   3.17    0       1       0       1       1.5     0
    −5.98    −0.26   4       0       1.42    3       0       1.83    0
    −4.1     1.59    5.83    0.83    0       .       0.33    1.17    .
    −6.92    −1.11   3.17    0       0       4.82    0.92    0       1.17
    −4.38    1.4     5.58    .       2.12    .       .       1.5     .
     . . .
    −4.05    1.81    6.33    2.83    0.5     .       0.12    0       .
    −3.06    2.65    7.17    0       4.42    .       0.25    0       .
    −6.13    −0.9    3.42    0.75    2.58    2.92    2.38    1.08    0
```

Example Code 2

There are 15 variables for each observation, which are age centered at 13 years old across three measures in columns 1 through 3, recoded binary variable of Using a Computer in columns 4 through 6, recoded continuous variable of Using a Computer in columns 7 through 9, recoded binary variable of Reading in columns 10 through 12, and recoded continuous variable of Reading in columns 13 through 15.

Variable:

```
Names = agec97 agec02 agec07 uCO97 uCO02 uCO07 mCO97 mCO02 mCO07
uRD97 uRD02 uRD07 mRD97 mRD02 mRD07;
Usevariables = agec97 agec02 agec07 uCO97 uCO02 uCO07 mCO97 mCO02 mCO07
uRD97 uRD02 uRD07 mRD97 mRD02 mRD07;
Categorical = uCO97 uCO02 uCO07 uRD97 uRD02 uRD07;
```

Input data:

```
−4.79   0.82   5.17   1   1   .   2.63   4.18   .      1   0   .   0.65   .      .
−6.82   −1.11  3.17   0   0   1   .      .      4.42   1   0   0   0.17   .      .
−6.79   −1.27  3.17   0   1   0   .      1      .      1   1   0   1      1.5    .
−5.98   −0.26  4      0   1   1   .      1.42   3      0   1   0   .      1.83   .
−4.1    1.59   5.83   1   0   .   0.83   .      .      1   1   .   0.33   1.17   .
−6.92   −1.11  3.17   0   0   1   .      .      4.82   1   0   1   0.92   .      1.17
−4.38   1.4    5.58   .   1   .   .      2.12   .      .   1   .   .      1.5    .
. . .
−4.05   1.81   6.33   1   1   .   2.83   0.5    .      1   0   .   0.12   .      .
−3.06   2.65   7.17   0   1   .   .      4.42   .      1   0   .   0.25   .      .
−6.13   −0.9   3.42   1   1   1   0.75   2.58   2.92   1   1   0   2.38   1.08   .
```