Psychological Methods
2009, Vol. 14, No. 3, 183–201

1082-989X/09/$12.00 DOI: 10.1037/a0015858

Evaluating Model Fit for Growth Curve Models: Integration of Fit Indices From SEM and MLM Frameworks

Wei Wu
University of Kansas

Stephen G. West
Arizona State University

Aaron B. Taylor
Texas A&M University

Evaluating overall model fit for growth curve models involves 3 challenging issues. (a) Three types of longitudinal data with different implications for model fit may be distinguished: balanced on time with complete data, balanced on time with data missing at random, and unbalanced on time. (b) Traditional work on fit from the structural equation modeling (SEM) perspective has focused only on the covariance structure, but growth curve models have four potential sources of misspecification: within-individual covariance matrix, between-individuals covariance matrix, marginal mean structure, and conditional mean structure. (c) Growth curve models can be estimated in both the SEM and multilevel modeling (MLM) frameworks; these have different emphases for the evaluation of model fit. In this article, the authors discuss the challenges presented by these 3 issues in the calculation and interpretation of SEM- and MLM-based fit indices for growth curve models and conclude by identifying some lines for future research.

Keywords: growth curve modeling, longitudinal data, multilevel modeling, structural equation modeling, model fit

Supplemental materials: http://dx.doi.org/10.1037/a0015858.supp

Growth curve modeling (GCM) has developed into one of the more important analytic approaches in the behavioral sciences. It has been widely applied in many areas of psychology, including clinical, developmental, educational, learning and memory, and personality. GCM can be characterized as an efficient method that simultaneously estimates intraindividual growth trajectories (represented by growth parameters, e.g., intercept and slope for linear growth) and interindividual differences in those growth parameters (Bryk & Raudenbush, 1987; Singer & Willett, 2003). GCM requires repeated observations using the same or equated measures on individuals. GCM enables investigators to predict future development, to study interindividual variation in growth trajectories, and to examine whether background characteristics and experimental treatments can account for variance in individual growth trajectories (Bryk & Raudenbush, 1987).

Modern GCM has largely developed out of three traditions in different disciplines. In psychology, latent curve analysis (e.g., Meredith & Tisak, 1990; Willett & Sayer, 1994; see also Tucker, 1958) has developed within the framework of structural equation modeling (SEM). Statistics, biostatistics, and econometrics have emphasized random coefficient regression models for longitudinal data (e.g., Laird & Ware, 1982; Rao, 1965). Education has emphasized the development of multilevel modeling (MLM) and its application to longitudinal data (e.g., Bryk & Raudenbush, 1987; Goldstein, 2003). These three approaches are closely related, but the advantages, disadvantages, and compatibility between the three approaches are still being explored (e.g., Chou, Bentler, & Pentz, 1998; Curran & Peterman, 2005; Mehta & West, 2000).

A central issue in GCM is the evaluation of the adequacy of the models. In practice, the SEM tradition has emphasized the use of chi-square goodness-of-fit tests and practical fit indices that evaluate the overall fit of the hypothesized model to the data. In contrast, the random coefficient regression models and MLM traditions, which are treated as

Wei Wu, Department of Psychology, University of Kansas; Stephen G. West, Department of Psychology, Arizona State University; Aaron B. Taylor, Department of Psychology, Texas A&M University.

Stephen G. West was supported by a sabbatical leave from Arizona State University at the Arbeitsbereich Methoden und Evaluation, Freie Universität Berlin (Germany), during the writing of this article.

Correspondence concerning this article should be addressed to Wei Wu, Department of Psychology, University of Kansas, Lawrence, KS 66045-7556. E-mail: wwei@ku.edu

identical here,[1] have emphasized the use of correlational measures of overall fit that evaluate the agreement between the estimated and observed responses. Fit indices offered by the SEM and MLM frameworks complement each other to reflect different sources of misfit in GCMs.

In this article, we focus on issues that arise in the evaluation of fit of GCMs from the perspectives of both the SEM and MLM frameworks. We begin the article with a consideration of three general types of longitudinal data structures that, as we see later, have implications for the calculation of measures of fit. We then briefly review critical features of the SEM and random coefficient/MLM approaches, including model specification, assumptions, and estimation procedures, highlighting aspects of estimation that have a close relationship with fit indices. We then introduce four sources of misfit in GCMs—two related to the mean structure and two related to the covariance structure. We show how they can be reflected in fit indices from the SEM and MLM frameworks, followed by a discussion of the factors that might affect the performance of those fit indices. Finally, we discuss issues and challenges in evaluating model fit for GCMs, which are introduced by the type of longitudinal data and the fact that GCMs involve different sources of misspecification from both mean and covariance structures. We summarize what is currently known and not known about assessing the fit of growth models and identify areas for future research. Space limitations preclude consideration of important issues related to hypothesis tests of parameter estimates and residual analysis (Bollen & Curran, 2006; Fitzmaurice, Laird, & Ware, 2004; Singer & Willett, 2003; Weiss, 2005).

Types of Longitudinal Data

Raudenbush (2001), building on earlier work by Ware (1985), proposed a classification in which three types of longitudinal data are distinguished (see Table 1).

Type I: Balanced on Time With Complete Data

Every person is observed at the same fixed set of time points, a design that is often termed a panel study. Suppose a research group wants to study the change of students' school achievement over Grades 1–3 (see Table 1, column 1). They repeatedly measure students' school achievement at the end of each grade from Grades 1 to 3. Here, grade in school is used to define the time scale. Every student was measured successfully, and there were no missing data. This structure is termed *Type I longitudinal data*.

Type II: Balanced on Time With Data Missing at Random (MAR)

The data collection plan *conceptually* specifies that every person should be observed at the same set of time points. However, in the actual data, some observations are MAR[2] (Little & Rubin, 2002; Rubin, 1976). Time must be discrete (observations taken at fixed time points), and the number of missing data patterns has to be limited so that all variances and covariances can be estimated. Data that are MAR may be an assumption of the researcher or may be built into the study design as in planned missingness designs (Graham, Taylor, Cumsille, & Olchowski, 2006). Table 1, column 2, illustrates this design. Student Case 1 was measured at all three time points. However, there was one observation missing at one of the three time points for Student Cases 2, 3, and 4. When all systematic sources of missingness in a panel study are related to measured study variables, this structure is termed *Type II longitudinal data*.

Type III: Unbalanced on Time

The design is unbalanced, which means every person is observed at a potentially different set of time points. The data become sparse (in the limit only 1 participant may be measured at a given time point) so that the variances and covariances cannot be estimated. Longitudinal data collection with a continuous (typically treated as random) time variable is an example of this type of data (see Table 1, column 3). Suppose age in months is used to define the time scale. Individuals are repeatedly measured on three occasions but at different sets of ages—for example, Case 1 was measured at ages 76.8, 79.2, and 81.3, whereas Case 2 was measured at ages 78.0, 79.5, and 80.1. This structure is termed *Type III longitudinal data.*

In practice, there are a large number of options for the time scale. The units may vary: For example, age could be expressed in years or months.[3] The origin (0-point) could be defined, for example, by birthdate, the date of entry into Grade 1, or the date of the last measurement. Each time scale has

[1] Although they developed separately, the random coefficient and MLM approaches are treated as identical in this article because they do not lead to models that have important distinctions in the present context.

[2] There are theoretically three types of missingness (Little & Rubin, 2002; Schafer & Graham, 2002). Data are missing completely at random (MCAR) when the probability of an observation being missing does not depend on any observed or unobserved measurements. Data are MAR when the probability of an observation being missing depends only on *observed* measurements. Finally, data are missing not at random (MNAR) when the probability of an observation being missing depends on its unobserved level. When data are MNAR, no method will produce unbiased estimates of the growth trend.

[3] In some cases, a Type III structure can be degraded into a Type I or Type II data structure. For example, if each child was measured at a random time point between the beginning and end of each school grade, the design could be analyzed using grade instead of age or exact time since the beginning of the study. However, using grade instead of exact time would involve considerable loss of information and introduce error into the measure of the time variable.

Table 1
Different Types of Longitudinal Data

Type I				Type II				Type III											
Time (grade)				Time (grade)				Time (age in months)											
ID	1	2	3	ID	1	2	3	ID	76.8	78.0	79.2	79.5	80.1	80.2	81.3	81.8	82.5	83.8	84.6	85.7
1	×	×	×	1	×	×	×	1	×		×				×					
2	×	×	×	2		×	×	2		×		×	×							
3	×	×	×	3	×	×		3						×			×		×	
4	×	×	×	4	×		×	4								×		×		×

Note. Type I is balanced on time with compete data. Type II is balanced on time with missing data assumed to be missing at random. Type III is unbalanced on time; time is continuous; and different numbers of observations may be collected for each participant. ID = identification number.

different implications for the interpretation of the results (for a discussion, see Biesanz, West, & Kwok, 2003). Types I and II are associated with many replications at each measurement point and a limited (if any) number of patterns of missing data. In contrast, Type III is associated with few (if any) replications at any time point and a large number of patterns of missing data.

The type of longitudinal data available is important for at least three reasons. First, SEM and MLM differ in their capability to analyze different types of longitudinal data. Second, different estimation methods are used for different types of longitudinal data. Third, the availability and interpretation of measures of model fit depend on the type of longitudinal data being analyzed.

The MLM and SEM Approaches to Modeling Growth: A Brief Review

MLM

MLM is a statistical technique that addresses clustered data and is typically used when participants are nested within groups or observations are nested within individuals. MLM takes into account the dependency between observations that is ignored by conventional multiple regression analysis (Cohen, Cohen, West, & Aiken, 2003). In longitudinal data, measurement occasions are nested within individuals, so MLM provides a method of estimating GCMs that yields correct standard errors for parameter estimates and, thus, correct significance tests and confidence intervals (MacCallum, Kim, Malarkey, & Kiecolt-Glaser, 1997).

For most applications, two-level models are adequate to represent growth. We now use a linear GCM with four time points to illustrate model specification for GCMs in a MLM framework (see Equations 1 and 2). All terms are defined below Equation 2. More complex forms of growth (e.g., quadratic) may be specified, but we consider only the basic linear GCM for simplicity.

$$\text{Level 1: } Y_{ij} = \pi_{0i} + \pi_{1i} Time_{ij} + \varepsilon_{ij},\ \varepsilon_{ij} \sim N(0, \sigma^2). \quad (1)$$

$$\text{Level 2: } \pi_{0i} = \gamma_{00} + \varsigma_{0i},\ \pi_{1i} = \gamma_{10} + \varsigma_{1i}, \quad (2)$$

$$\begin{bmatrix} \varsigma_{0i} \\ \varsigma_{1i} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mathbf{G} = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix}\right),$$

where *MVN* = multivariate normally distributed; *i* indicates individual, $i = 1, \ldots, N$; and *j* indicates time point. In the example, $j = 1, 2, 3, 4$.

The Level-1 submodel (Equation 1) is specified to capture the shape of the within-individual growth trends by predicting the response variable (Y_{ij}) as a function of time ($Time_{ij}$). The parameters determining the individual growth trajectories are called growth parameters, which for a linear growth trajectory are the intercept (π_{0i}) and slope (π_{1i}). The Level-1 model also includes residuals (ε_{ij}) that represent deviations of the observed data from the predicted individual growth curves. Conventionally in MLM, ε_{ij} are assumed to follow a normal distribution with a mean of 0, and a covariance matrix (called the **R** matrix) with a constant variance σ^2 and no covariance over time. For growth curve models, the assumptions of constant variance and no covariance of residuals over time may be unrealistic.[4]

The basic Level-2 submodel (Equation 2) is specified to capture the interindividual differences in growth parameters. At this level, the growth parameters for the intercept (π_{0i}) and slope (π_{1i}) conceptually become outcome variables, and predictors that might account for their variation across individuals can be entered into the model. In the basic growth curve model in which no predictor is specified, π_{0i} and π_{1i} are represented by their predicted population

[4] There are three different traditions with respect to residuals in GCMs: (a) MLM has used the default that error variances are constrained equal over time; (b) some researchers (e.g., Browne & Du Toit, 1991) have advocated using constraints that allow the error variances to increase following an ordinally increasing structure; and (c) SEM has used the default that error variances are freely estimated over time. The decrease in restrictions as the models move from (a) to (c) can have implications for the fit of the models. Depending on the data and the model, the restrictions can in some cases lead to appreciable changes in model fit.

mean intercept (γ_{00}) and mean slope (γ_{10}). γ_{00} and γ_{10} are constant across individuals and are termed fixed effects. Deviations of individual growth parameters from the population means are ς_{0i} and ς_{1i} for the intercept and slope, respectively. ς_{0i} and ς_{1i} vary across individuals and are termed random effects. **G** is the covariance matrix of the random effects, which is also called the between-individuals covariance matrix. τ_{00} is the variance of the intercepts, τ_{11} is the variance of the slopes, and $\tau_{01}(=\tau_{10})$ is the covariance of the intercepts and slopes.

For some applications, it is useful to combine the two levels together to obtain a mathematically equivalent reduced form or mixed model:

$$Y_{ij} = \gamma_{00} + \gamma_{10} Time_{ij} + \varsigma_{0i} + \varsigma_{1i} Time_{ij} + \varepsilon_{ij}. \quad (3)$$

Equation 4 below expresses the model in matrix form. The key insight from Equation 4 is that there is a term $\mathbf{X}_i\Gamma$ that represents the fixed effects (mean parameter estimates in the sample), a term $\mathbf{Z}_i\mathbf{b}_i$ that represents the random effects (summarizes the individual deviations from the mean levels), and a random error term $\mathbf{e}_i$.

$$\mathbf{Y}_i = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix} = \begin{bmatrix} 1 & Time_{i1} \\ 1 & Time_{i2} \\ 1 & Time_{i3} \\ 1 & Time_{i4} \end{bmatrix} \times \begin{bmatrix} \gamma_{00} \\ \gamma_{10} \end{bmatrix} + \begin{bmatrix} 1 & Time_{i1} \\ 1 & Time_{i2} \\ 1 & Time_{i3} \\ 1 & Time_{i4} \end{bmatrix} \times \begin{bmatrix} \varsigma_{0i} \\ \varsigma_{1i} \end{bmatrix} + \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4} \end{bmatrix}$$

$$\mathbf{Y}_i = \mathbf{X}_i\Gamma + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i. \quad (4)$$

Here, Γ is the vector of fixed effects; $\mathbf{b}_i$ is the vector of random effects for individual i; $\mathbf{X}_i$ is the design matrix for fixed effects; and $\mathbf{Z}_i$ is the design matrix for random effects. $\mathbf{Z}_i$ can be equal to $\mathbf{X}_i$ or a submatrix of $\mathbf{X}_i$. As can be seen in Equation 4, it is easy to accommodate random time for each individual (each individual may have a different number of repeated measurements and be measured at different occasions) in the MLM framework.

Model implied marginal means. Using the fixed effects, we can calculate the model implied mean response profile, termed *marginal means* (Equation 5).

$$\hat{\boldsymbol{\mu}}_i = E(\mathbf{Y}_i) = \mathbf{X}_i\hat{\Gamma}. \quad (5)$$

The estimated marginal means are the overall mean level predicted by the growth model at each measurement wave for the sample of participants.

Model implied covariance matrix. Combining the **G** and **R** matrices using Equation 6, we can calculate the estimated model implied covariance matrix.

$$\hat{\boldsymbol{\Sigma}}_i = COV(\mathbf{Y}_i) = \mathbf{Z}_i\hat{\mathbf{G}}\mathbf{Z}_i' + \hat{\mathbf{R}}_i, \quad (6)$$

where $\hat{\mathbf{G}}$ is the model implied between-individuals covariance matrix, and $\hat{\mathbf{R}}_i$ is the model implied within-individual covariance matrix. The values of $\hat{\boldsymbol{\Sigma}}_i$ represent the variances at each measurement wave and the covariances between measurement waves for each individual that are predicted by the growth model.

In the basic growth curve model for Type I data, we can drop the i subscripts. All individuals are measured at the same set of common times so there is only one predicted mean vector $\hat{\boldsymbol{\mu}}$ for the set of measurement occasions. Similarly, $\hat{\boldsymbol{\Sigma}}_i$ can be simplified to $\hat{\boldsymbol{\Sigma}}$ because both the $\mathbf{Z}_i\hat{\mathbf{G}}\mathbf{Z}_i'$ and $\hat{\mathbf{R}}_i$ matrices are constant across individuals.

Model implied conditional means. In MLM, we can estimate the random effects ($\mathbf{b}_i$) for each individual using the estimated best linear unbiased predictor (EBLUP; see Equation 7 below). Then, the estimated random effects can be used to calculate the estimated individual response profile, $\hat{\mathbf{Y}}_i$ (see Equation 8). In other words, we use information from the mean growth line and the individual i's deviation from the mean slope and intercept to estimate a growth line specifically characterizing individual i. $\hat{\mathbf{Y}}_i$ is also called the vector of *conditional means* because it is conditional on the random effects for individual i ($\mathbf{b}_i$).

$$\hat{\mathbf{b}}_i = \hat{\mathbf{G}}\mathbf{Z}_i'\hat{\boldsymbol{\Sigma}}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\hat{\Gamma}), \quad (7)$$

$$\hat{\mathbf{Y}}_i = E(\mathbf{Y}_i|\mathbf{b}_i) = \mathbf{X}_i\hat{\Gamma} + \mathbf{Z}_i\hat{\mathbf{b}}_i = (\hat{\mathbf{R}}_i\hat{\boldsymbol{\Sigma}}_i^{-1})\mathbf{X}_i\hat{\Gamma} + \mathbf{Z}_i\hat{\mathbf{G}}\mathbf{Z}_i'\hat{\boldsymbol{\Sigma}}_i^{-1}\mathbf{Y}_i, \quad (8)$$

where again $\mathbf{X}_i\hat{\Gamma}$ is the vector of model implied marginal means for individual i (Fitzmaurice et al., 2004).

A careful look at Equation 8 shows that EBLUP shrinks the *ith* individual's predicted response profile toward the mean response profile based on the entire sample ($\mathbf{X}_i\hat{\Gamma}$). The shrinkage depends on the relative magnitude of the within-individual variability ($\hat{\mathbf{R}}_i$) and the between-individuals variability ($\mathbf{Z}_i\hat{\mathbf{G}}\mathbf{Z}_i'$). When the within-individual variability is large relative to the between-individuals variability, more weight is assigned to the estimated marginal means ($\mathbf{X}_i\hat{\Gamma}$) than to the *ith* individual's observed responses ($\mathbf{Y}_i$) (Verbeke & Molenberghs, 2000).

SEM

SEM is a comprehensive statistical technique used to test hypotheses about relations among observed and latent variables (Hoyle, 1995). Meredith and Tisak (1990) showed that SEM can be used to estimate growth curve models if the growth parameters are treated as latent variables and repeated measures as multiple indicators of the latent variables.

Figure 1 depicts a linear GCM in the SEM framework, which is identical to the linear GCM represented earlier in the MLM framework. The factor loadings associated with the intercept are all fixed at 1 because the intercept is invariant across time. The loadings associated with the slope are usually fixed at values proportional to the time of each measurement occasion (t_1, t_2, t_3, and t_4 in Figure 1). For example, if data were collected at baseline (0 months), 6 months, 12 months, and 24 months, one possible scaling of these loadings would be $t_1 = 0$, $t_2 = 6$, $t_3 = 12$, and $t_4 = 24$; another would be $t_1 = 0$, $t_2 = 1$, $t_3 = 2$, and $t_4 = 4$. Here, the matrix of factor loadings is identical to the matrix of $\mathbf{X}_i$ and $\mathbf{Z}_i$ in the MLM example. The mean of the latent variables estimates the population mean of the growth parameters (γ_{00} and γ_{10}, fixed effects). The covariance matrix of the latent variables is identical to the **G** matrix in the MLM framework. The covariance matrix of the unique factors is identical to the **R** matrix in the MLM framework.

Basic Assumptions

Both frameworks usually assume that the longitudinal responses, Level-1 residuals, and Level-2 random effects have a multivariate normal distribution. They also assume that the Level-2 random effects are independent of the Level-1 residuals, and the Level-1 residuals are usually assumed to be independent of one another. Moderate departures from normality do not severely affect estimation of the fixed effects, although a correction may be needed to produce accurate standard errors (Verbeke & Molenberghs, 2000). Violations of the independence assumptions are more problematic, leading to biased test statistics (Satorra, 1992).

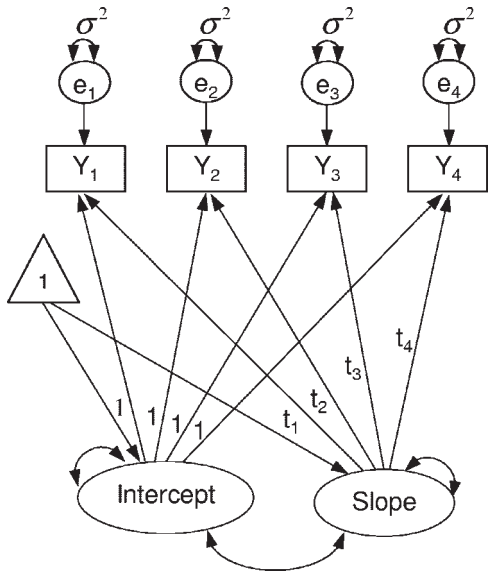

Figure 1. A linear latent growth curve model. Y_1 to Y_4 are four waves of repeated measures. Intercept and slope are the two growth parameters that are correlated with each other. The factor loadings are fixed at 1 for the intercept factor and are fixed at time at each wave (t_1 to t_4) for the slope factor. The two paths pointing from the triangle to the two factors represent the means of the intercept and slope factors; e_1 to e_4 are residuals (Level-1 residuals) for Y_1 to Y_4, respectively. The residual variances (or unique variances) are constrained to be equal so the structural equation modeling model is identical to the multilevel modeling model presented earlier. In the model portrayed, the triangle indicates that the means of intercept and slope are estimated.

Comparison of MLM and SEM in GCM

Although MLM and SEM represent the growth model in different ways, they share the same basic rationale when modeling growth. They also yield similar results across a wide range of models, including all linear growth models as well as some nonlinear growth models (Chou et al., 1998; MacCallum et al., 1997; Mehta & West, 2000; B. O. Muthén & Curran, 1997).

However, associated with the different methodologies they use, both MLM and SEM have unique advantages and limitations. SEM has advantages over MLM in flexibility in modeling data from Types I and II longitudinal data. For instance, SEM can estimate GCMs in which loadings of measured variables on the growth parameters are freely estimated; can easily freely estimate the elements in the **R** and **G** matrix; can model a latent outcome variable with multiple indicators at each time point; and can include other variables that can serve as correlates, predictors, or consequences (outcomes) of the latent growth parameters (MacCallum et al., 1997; Meredith & Tisak, 1990; B. O. Muthén & Curran, 1997).

On the other hand, MLM allows for simpler model specification and is more efficient computationally in yielding results. MLM is also better at incorporating additional levels of clustering (e.g., repeated measures on individuals clustered within groups), for which the SEM approach quickly becomes unwieldy (Mehta & West, 2000; B. O. Muthén & Curran, 1997). Of particular importance in the present context, MLM can directly handle all three types of longitudinal data described earlier. Traditionally, SEM has been able to handle only the first two types of longitudinal data because it assumes that a common covariance matrix exists across individuals, which is only plausible for Type I and Type II longitudinal data structures. The introduction of full information maximum likelihood (FIML) estimation procedures in several software packages now allows SEM to be used to estimate models with a wider variety of data structures, including Type III longitudinal data. However, standard SEM fit indices *cannot* be calculated for this latter type of data.

Estimators for GCMs

Understanding the functions used to estimate GCMs is important because these estimators are the basis for chi-

square tests of overall model fit and ultimately for the calculation of practical fit indices for Type I and Type II longitudinal data that are discussed later (Hu & Bentler, 1998). We focus on two estimators that are commonly used in estimating GCMs: FIML and standard maximum likelihood (ML).

For FIML, the log-likelihood of a set of sample data $\mathbf{Y}$ given a vector of parameters $\boldsymbol{\theta}$ can be expressed as follows.

$$\ln L_{FIML}(\mathbf{Y}|\boldsymbol{\theta}) = -\frac{1}{2}\sum_{i=1}^{N} n_i \log(2\pi) - \frac{1}{2}\sum_{i=1}^{N} n_i \log|\hat{\boldsymbol{\Sigma}}_i(\boldsymbol{\theta})| - \frac{1}{2}\sum_{i=1}^{n} ((\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i(\boldsymbol{\theta}))' \hat{\boldsymbol{\Sigma}}_i(\boldsymbol{\theta})^{-1} (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i(\boldsymbol{\theta})). \quad (9)$$

Here, $\boldsymbol{\theta}$ is the vector of parameters, N is the number of individuals, n_i is the number of observations for individual i, $\hat{\boldsymbol{\mu}}_i(\boldsymbol{\theta})$ is the model implied marginal mean vector, and $\hat{\boldsymbol{\Sigma}}_i(\boldsymbol{\theta})$ is the model implied covariance matrix for individual i. The third term of Equation 9, $\frac{1}{2}\sum_{i=1}^{n} ((\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i(\boldsymbol{\theta}))' \hat{\boldsymbol{\Sigma}}_i(\boldsymbol{\theta})^{-1} (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i(\boldsymbol{\theta}))$, reflecting the fit of the mean structure, is of particular importance in GCMs; its effects have received relatively little study. In addition to multivariate normality, FIML assumes that data are MAR (Rubin, 1976). FIML takes advantage of all the available observed data for each individual. FIML can be applied to all three types of longitudinal data. For GCMs with Type I longitudinal data, the i subscript in Equation 4 can be dropped, and FIML simplifies to ML. Equation 10 shows the log-likelihood function for ML. Equation 11 is the discrepancy function for ML. Maximizing Equation 10 and minimizing Equation 11 will lead to equivalent parameter estimates. ML can be also used with Type II longitudinal data through the use of a multiple-group approach in which equality constraints are set across groups having different patterns of missing data, so long as a sufficient number of cases follow each pattern of missing data (B. O. Muthén, Kaplan, & Hollis, 1987).

$$\ln L_{ML}(\mathbf{Y}|\boldsymbol{\theta}) = -\frac{1}{2}\sum_{i=1}^{N} n \log(2\pi) - \frac{N}{2}\log|\hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta})| - \frac{1}{2}\sum_{i=1}^{n} ((\mathbf{Y}_i - \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}))' \hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta})^{-1} (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}(\boldsymbol{\theta})), \quad (10)$$

$$F_{ML}(\boldsymbol{\theta}) = (\bar{\mathbf{Y}} - \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}))' \hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta})^{-1} (\bar{\mathbf{Y}} - \hat{\boldsymbol{\mu}}(\boldsymbol{\theta})) + \ln|\hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta})| - \ln|\mathbf{S}| + tr\hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta})^{-1}\mathbf{S} - p, \quad (11)$$

where $\bar{\mathbf{Y}}$ is the sample mean vector, $\mathbf{S}$ is the sample covariance matrix, and p is the number of nonduplicated elements in the covariance matrix.

The Sources of Misfit in GCM

To evaluate model fit for GCM, it is necessary to understand the sources of misfit. Because GCMs involve both mean and covariance structures, the misfit can come from the mean, covariance, or both structures. Figure 2 presents a hierarchical depiction of the sources of misfit for GCMs. At the highest level of the hierarchy is the GCM. The GCM can be partitioned into the mean structure and covariance structure. The mean structure can be further partitioned into marginal and conditional mean structures. The covariance structure, which we call the total covariance structure, can be further partitioned into within-individual and between-individuals covariance matrices ($\mathbf{R}$ and $\mathbf{G}$ matrices). The total covariance structure is a function of the $\mathbf{R}$ and $\mathbf{G}$ matrices (see Equation 6). Potential sources of misfit exist in each of the four elements at the lowest level of the hierarchy. We initially describe each of these four elements below and then describe their interrelationships.

*Misfit in the Within-Individual Covariance Matrix (**R** Matrix)*

The within-individual covariance matrix contains the variances and covariances of the Level-1 residuals. Some elements in the $\mathbf{R}$ matrix might not be correctly specified. For example, in the basic MLM framework, the variance of the Level-1 residuals is usually assumed to be constant over time (see Footnote 4), and there is no covariance between Level-1 residuals from different measurement waves. In practice, the variances might differ at different time points. In addition, there might be covariances among the Level-1 residuals that cannot be accounted for by the variables included in the Level-1 model (e.g., time).

*Misfit in the Between-Individuals Covariance Matrix (**G** Matrix)*

The between-individuals covariance matrix contains the variance and covariances of the growth parameters. Some

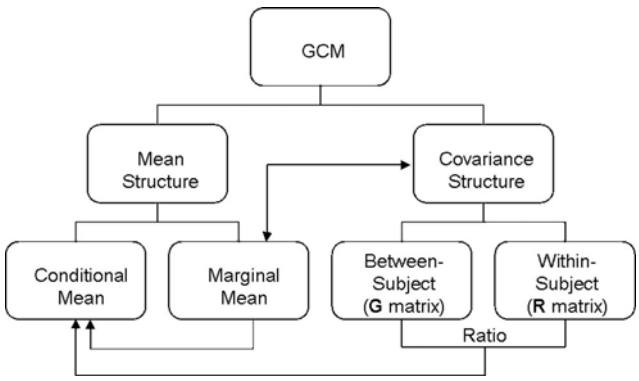

Figure 2. Sources of misfit in growth curve modeling (GCM) models.

elements in the matrix might not be correctly specified. For example, the variances or covariances of growth parameters might be wrongly constrained to be 0.

Misfit in the Marginal Mean Structure

The marginal mean structure is the structure (functional form) for the mean responses at each of the measurement occasions across individuals. An incorrect functional form might be specified for the mean growth trajectory. As a result, the model implied marginal means ($E(\mathbf{Y}_i)$; see Equation 5) would not agree with the sample means ($\bar{\mathbf{Y}}_i$).

Misfit in the Conditional Mean Structure

The conditional mean structure is the structure (functional form) for the responses at each of the measurement occasions for each individual. An incorrect functional form might be specified for the individual growth trajectories. As a result, the model estimated responses at each measurement wave for individuals ($E(\mathbf{Y}_i|\mathbf{b}_i)$; see Equation 8) might not agree with their observed individual responses ($\mathbf{Y}_i$).

Relationships Among the Sources of Misfit

Relationships between the fit of the marginal mean structure and the fit of the covariance structures. The fit of the covariance structures will be affected by the discrepancy between the observed and estimated marginal means. The total covariance structure is calculated on the basis of the residuals from the marginal means so that the magnitude of the discrepancy in the covariance structure will increase as the marginal means become increasingly misspecified. In contrast, marginal means are a function only of the fixed effects as shown in Equation 5. The estimates of the marginal means are far less sensitive to specification of the covariance structure; indeed, for linear models, the marginal means are asymptotically independent of the covariance structure for Type I longitudinal data when the standard assumptions are met (Verbeke & Lesaffre, 1997; Yuan & Bentler, 2004). However, for realistic sample sizes, the fit of the marginal mean structure will still be affected by the estimated covariance structure because the residuals in means are weighted by the estimated covariance structure (see Equations 9, 10, and 11).

Relationships between the fit of the conditional mean structure and the fit of the marginal mean and covariance structures. Recall that the fit of the conditional mean structure (fit of the individual functional form) measures the agreement between the observed and estimated individual responses. The fit of the conditional mean structure is dependent on both the fit of the marginal mean structure and the fit of the covariance structure because the estimated individual responses ($E(\mathbf{Y}_i|\mathbf{b}_i)$) are a function of both the marginal mean and covariance structures (specifically, the ratio of the between- to within-subject variability [see Equation 8]). Knowing either the fit of the marginal mean or the fit of the covariance structure alone is *not* sufficient to make an inference about the fit of the conditional mean structure.

The fit of the marginal mean structure (functional form for the average growth trajectory) alone cannot guarantee the fit of the conditional mean structure (functional form for individual growth trajectories) because the functional form for the average growth trajectory and individual growth trajectories can be inconsistent (Singer & Willet, 2003). Figure 3 shows an example in which the individual growth curves have a quadratic form, but the average growth trajectory is linear. In practice, we should choose the functional form that fits most of the individual growth trajectories instead of just the average mean trajectory. Thus, by checking only the fit of the marginal mean structure, one may end up with an incorrect functional form for the GCM.

The fit of the covariance structure alone cannot guarantee the fit of the conditional mean structure either, because we can always improve the fit of the covariance structure by freeing more elements in the within-individual covariance matrix (**R** matrix; Marsh, Hau, & Wen, 2004). In this case, the total covariance structure (see Equation 5), which is a combination of both within-individual and between-individual covariance matrices, might show a good fit to the data. However, the true between-individuals and within-individual variability will not be correctly captured. For example, if the correct functional form were quadratic, but the data were fitted with a linear GCM, an adequate fit of the covariance structure might be achieved by allowing the residuals to be correlated or to vary over time—practices that can be easily done in SEM. However, the between-individuals variability would be underestimated and the within-individual variability would be overestimated.

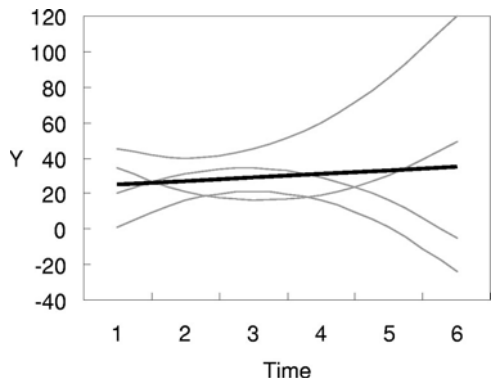

Figure 3. Individual and average growth trajectories: An example.

An interesting theoretical question is the extent to which the adequacy of both the marginal mean and the covariance structures justifies the adequacy of the conditional mean structure (the fit of the individual functional form). Under standard assumptions, if the marginal means, the within-individual, and the between-subjects variability were all adequately captured, then fit of the conditional mean structure would be achieved. However, as described above, even though the total covariance structure might provide an adequate fit to the data, it is difficult to know whether the within and between-subjects variability have been adequately captured. Given that both the marginal mean and covariance structures provide a good fit to the data, there might be two explanations for the misfit in the conditional mean structure: (a) the individual functional form is incorrectly specified; and (b) the individual functional form is correctly specified, but the ratio of between-individuals to within-individual variability is not correctly captured. The second explanation might be true when there is misspecification in the between-individuals covariance matrix, within-individual covariance matrix, or both. To rule out the second explanation, researchers need to specify the between-individuals and within-individual covariance structures carefully. Verbeke and Molenberghs (2000) recommended freeing as many elements as possible in the between-individuals covariance matrix first before considering the within-individual covariance matrix. In this way, one can know that both within-individual and between-individuals covariance matrices have been *optimally* specified given the fitted individual functional form. If one still detects the misfit in the conditional mean structure, then that misfit must be due to the first explanation: incorrect individual functional form. Clearly, more work needs to be done to understand the relationship between the fit of the marginal mean, conditional mean, and covariance structures.

Introduction to Fit Indices in SEM and MLM Frameworks

As described above, GCMs can be estimated in both the SEM and MLM frameworks. Both frameworks provide information on model fit. Compared with MLM, SEM provides more varieties of model fit indices. Though less familiar, the MLM-based fit indices have their own strengths, as described below. Our presentation starts with an introduction of the available fit indices in both frameworks. We then discuss what source(s) of misfit can be detected by these fit indices. Given space limitations, we do not attempt to consider every SEM fit index. Rather, we review basic dimensions that can be used to classify SEM-based fit indices and show an example fit index within each category. In SEM, researchers have conventionally been expected to evaluate the adequacy of the fitted model using fit indices. In MLM, global model fit has received far less emphasis, with researchers often only reporting the significance of parameter estimates.

SEM-Based Fit Indices

Yuan (2005) proposed that SEM fit indices can be classified into two major categories on the basis of whether the fit index is defined directly through a likelihood ratio test (T) or residuals in the mean and covariance matrices. Yuan argued that all of the fit indices can be treated as weighted functions of residuals, but the fit indices that are defined through test statistics utilize theoretically more optimal weight functions.

Sun (2005) proposed a more detailed classification of the SEM-based fit indices. He argued that three dimensions can be used to classify SEM-based fit indices.

1. Sample based versus population based. Sample-based fit indices measure the discrepancy between the observed and model implied mean vector and observed and model implied covariance matrices. The likelihood ratio test statistic (T) is the most popular sample-based fit index. Given multivariate normality,[5] T follows a chi-square distribution, and thus is often called the chi-square test statistic. Population-based fit indices estimate the discrepancy between the model implied covariance matrix and the population covariance matrix. Because the population covariance matrix is unknown, the noncentrality parameter (λ) is used to represent the population discrepancy. $T - df$ is a sample estimate of λ. λ is dependent on sample size; thus, researchers commonly rescale $T - df$ by dividing by $N - 1$, which leads to $d = (T - df)/(N - 1)$ (McDonald, 1989; Steiger, Shapiro, & Browne, 1985).

2. Absolute versus relative. Absolute fit indices evaluate the model fit of the hypothesized model without a comparison with a baseline model, whereas relative fit indices measure the specific improvement in model fit of the hypothesized model relative to a baseline model (Bollen & Curran, 2006). The baseline model is chosen to estimate as few parameters as are reasonable. Widaman and Thompson (2003) have noted the importance of specifying the correct baseline model to derive a valid relative fit index.

3. Adjustment versus no adjustment for model complexity. T will always decrease as model complexity increases. Model complexity is usually indicated by df or number of free parameters. Some fit indices impose an adjustment for model complexity; some do not. Three different strategies are used to impose a penalty for model complexity: linearly combining T with a weighted model df, dividing T by

[5] T will also follow a chi-square distribution under weaker assumptions, such as asymptotic independence (see Satorra, 1992).

the model *df*, or multiplying a relative fit index by a parsimony index.

The combination of the three dimensions yields eight categories. Table 2 shows examples of fit indices in each category and their basic properties.

MLM-Based Fit Indices

MLM can be seen as an extension of multiple regression analysis. R^2 is often used to reflect the fit of the multiple regression model. In OLS regression, R^2 measures the proportion of variation in the outcome variable that can be accounted for by the predictors in the regression model. R^2 is also equal to the squared correlation between the observed outcome and estimated outcome.

Pseudo R^2. Singer and Willet (2003, pp. 102–103) proposed a pseudo R^2 statistic to summarize the proportion of total outcome variability explained by predictors in the model. They defined pseudo $R^2 = r^2_{Y\hat{Y}}$, which is the squared Pearson correlation between the observed individual responses (Y) and the estimated marginal mean responses ($\hat{\bar{Y}}$). Pseudo R^2 reflects the agreement between the observed individual responses and estimated marginal means.

Conditional concordance correlation (CCC; Vonesh, Chinchilli, & Pu, 1996). The CCC was originally developed in biostatistics to assess the agreement between two continuous measures from different raters or methods (Lin, 1989). Vonesh et al. (1996) and Vonesh and Chinchilli (1997) extended the CCC to assess the agreement between the observed individual responses and conditional means (estimated individual responses) in mixed effects model settings. This conditional CCC is defined in Equation 12 (Vonesh et al., 1996):

Table 2
Classifications and Basic Properties of SEM-Based Fit Indices

Hierarchical classification (Sun, 2005)	Fit indices	Reference	Chi-square versus residual based (Yuan, 2005)	Direction	Normed (0–1)
Sample–absolute–unadjusted	T	Bollen (1989)	Chi-square	Small is good	No
	$SRMR^a = \sqrt{\sum_j \sum_k r^2_{jk}/p^*}$	Jöreskog and Sörbom (1981)	Residual	Small is good	No
	$WRMR^b = \sqrt{\sum_r^e \frac{(s_r - \hat{\sigma}_r)^2}{v_r}/e}$	L. K. Muthén and Muthén (1998–2007)	Residual	Small is good	No
	$GFI^c = 1 - T_h/\min[F(S;\Sigma(0))]$	Jöreskog and Sörbom (1984)	Chi-square	Large is good	Yes
Sample–absolute–adjusted	$AGFI = 1 - (1 - GFI)p^*/df_h$	Jöreskog and Sörbom (1984)	Chi-square	Large is good	No
Sample–relative–unadjusted	$NFI = (T_b - T_h)/T_b$	Bentler and Bonett (1980)	Chi-square	Large is good	Yes
Sample–relative–adjusted	$TLI = [(T_b/df_b) - (T_h/df_h)]/[(T_b/df_b) - 1]$	Tucker and Lewis (1973)	Chi-square	Large is good	No
Population–absolute–unadjusted	$Mc^d = \exp(-\frac{1}{2}d_h)$	McDonald (1989)	Chi-square	Large is good	No
Population–absolute–adjusted	$RMSEA = \sqrt{\max(d_h/df_h, 0)}$	Steiger and Lind (1980)	Chi-square	Small is good	No
Population–relative–unadjusted	$CFI = 1 - \max[d_h, 0]/\max[d_h, d_b, 0]$	Bentler (1990)	Chi-square	Large is good	Yes
Population–relative–adjusted	$PCFI = CFI \times df_h/df_b$	James et al. (1982)	Chi-square	Large is good	Yes

Note. *SEM* = structural equation modeling; T = likelihood ratio test statistic (chi-square test statistic); *SRMR* = standardized root-mean-square residual; *WRMR* = weighted root-mean-square residual; *GFI* = goodness-of-fit index; *AGFI* = adjusted goodness-of-fit index; *NFI* = normed fit index; *TLI* = Tucker–Lewis index; *Mc* = McDonald's measure of centrality; *RMSEA* = root-mean-square error of approximation; *CFI* = comparative fit index; *PCFI* = parsimony version of CFI; h = hypothesized model; b = baseline model.

[a] r_{jk} is a standardized residual from a covariance matrix with j rows and k columns; p^* is the number of nonduplicated elements in the covariance matrix. [b] s_r is an element of the sample statistics vector, including sample mean and covariance parameters; $\hat{\sigma}_r$ is the estimated model counterpart of s_r; v_r is an estimate of the asymptotic variance of s_r; and e is the number of sample mean and covariance parameters. [c] $\min[F(\mathbf{S};\mathbf{\Sigma}(\mathbf{0}))]$ is the minimum value of the discrepancy function with all elements in the population covariance matrix assumed to be 0. [d] $d = (T - df)/(N - 1)$, where $T - df$ is the sample estimate of the noncentrality parameter.

$$r_c = 1 - \frac{\sum_{i=1}^{n} (\mathbf{Y_i} - \hat{\mathbf{Y}}_\mathbf{i})'(\mathbf{Y_i} - \hat{\mathbf{Y}}_\mathbf{i})}{\left\{ \begin{array}{c} \sum_{i=1}^{n} (\mathbf{Y_i} - \bar{Y}\mathbf{1_i})'(\mathbf{Y_i} - \bar{Y}\mathbf{1_i}) + \sum_{i=1}^{n} (\hat{\mathbf{Y}}_\mathbf{i} - \hat{\bar{Y}}\mathbf{1_i})' \\ (\hat{\mathbf{Y}}_\mathbf{i} - \hat{\bar{Y}}\mathbf{1_i}) + \mathrm{N}(\bar{Y} - \hat{\bar{Y}})^2 \end{array} \right\}}, \quad (12)$$

where $\bar{Y}$ is the grand mean of the observed responses, $\hat{\bar{Y}}$ is the grand mean of the estimated responses, $\mathbf{1_i}$ is an identity vector that contains a column of 1s, $\mathbf{Y_i}$ is the vector of observed individual responses for individual i, and $\hat{\mathbf{Y}}_\mathbf{i}$ is the vector of estimated individual responses (conditional means) for individual i. As shown before, $\hat{\mathbf{Y}}_\mathbf{i}$ can be calculated using Equation 8 for the linear mixed model. N is the total number of observations; n is the total number of individuals. Note that this formula assumes homogeneity in $\mathbf{Y_i}$. If there is heterogeneity in $\mathbf{Y_i}$, $\mathbf{Y_i}$, and $\hat{\mathbf{Y}}_\mathbf{i}$ need to be weighted by $\mathbf{\Sigma}_i^{-1/2}$, where $\mathbf{\Sigma}_i$ is the covariance matrix of $\mathbf{Y_i}$.

Note that the numerator of the second term in Equation 12, $\sum_{i=1}^{n} (\mathbf{Y_i} - \hat{\mathbf{Y}}_\mathbf{i})'(\mathbf{Y_i} - \hat{\mathbf{Y}}_\mathbf{i})$, is the expected square of the deviation of the pairs of $\mathbf{Y_i}$ and $\hat{\mathbf{Y}}_\mathbf{i}$ from the 45° line through the origin (perfect agreement between $\mathbf{Y_i}$ and $\hat{\mathbf{Y}}_\mathbf{i}$). The denominator is the expected squared deviation of the pairs of $\mathbf{Y_i}$ and $\hat{\mathbf{Y}}_\mathbf{i}$ from the 45° line when $\mathbf{Y_i}$ and $\hat{\mathbf{Y}}_\mathbf{i}$ are not correlated. The lower the deviation of the pairs of $\mathbf{Y_i}$ and $\hat{\mathbf{Y}}_\mathbf{i}$ from the 45° line, the higher the conditional CCC will be. The conditional CCC can be used to test the fit of the conditional mean structure.

Average CCC (Vonesh et al., 1996). If we use the estimated marginal means ($\hat{\tilde{\mathbf{Y}}}_\mathbf{i}$) to replace $\hat{\mathbf{Y}}_\mathbf{i}$ in Equation 12, then we can use the CCC to evaluate the agreement between the observed individual responses and the estimated marginal mean responses. Vonesh et al. (1996) referred to this CCC as *average CCC*.

$$r_c = 1 - \frac{\sum_{i=1}^{n} (\mathbf{Y_i} - \hat{\tilde{\mathbf{Y}}}_\mathbf{i})'(\mathbf{Y_i} - \hat{\tilde{\mathbf{Y}}}_\mathbf{i})}{\left\{ \begin{array}{c} \sum_{i=1}^{n} (\mathbf{Y_i} - \bar{Y}\mathbf{1_i})'(\mathbf{Y_i} - \bar{Y}\mathbf{1_i}) + \sum_{i=1}^{n} (\hat{\mathbf{Y}}_\mathbf{i} - \hat{\bar{Y}}\mathbf{1_i})' \\ (\hat{\mathbf{Y}}_\mathbf{i} - \hat{\bar{Y}}\mathbf{1_i}) + \mathrm{N}(\bar{Y} - \hat{\bar{Y}})^2 \end{array} \right\}} \quad (13)$$

The distribution of the CCC measures is asymptotically normal (Lin, 1989). The standard Fisher r to z transformation can be used to improve the normal approximation of the CCC measures and to calculate the confidence interval for the CCC measures (see Cohen et al., 2003).

Attractive features of CCC measures. According to Vonesh et al. (1996), the CCC family of measures has several advantages.

1. They have an intuitively reasonable interpretation. They directly measure the level of agreement between the observed responses $\mathbf{Y}_i$ and conditional or marginal model implied responses ($\hat{\mathbf{Y}}_i$ or $\hat{\tilde{\mathbf{Y}}}_i$).

2. The 45° line of identity where $\mathbf{Y}_i = \hat{\mathbf{Y}}_i$ or $\mathbf{Y}_i = \hat{\tilde{\mathbf{Y}}}_i$ serves as a point of reference for the CCC measures indicating perfect fit.

3. CCC measures have well defined endpoints (−1, 1). A value of 1 indicates a perfect fit, and a value smaller than 0 indicates a lack of fit. A null model need not be specified.

4. The variability of the points around the best fitting straight line measures how far each observation deviates from the optimal individual or mean trajectory fit to the data (precision). The discrepancy in slopes between the best fitting line and the 45° line (observed responses vs. estimated responses) indicates model accuracy. Note that the Pearson correlation can measure the degree of precision but not the degree of accuracy.

5. CCC measures are semiparametric coefficients that do not require specification of a likelihood function. They may be robust to nonnormal distributions.

Figure 4 is a simplified illustration of several of these properties. It shows the observed individual responses for five measurement occasions for only a single participant rather than the full sample. The measure of accuracy is the angle between the 45° line and the best fitting line. The spread of the observed responses around the best fitting line provides a measure of precision.

Fit Indices for Different Sources of Misfit

Fit indices for covariance structure. SEM-based fit indices were originally developed to detect misfit in the covariance structure. Numerous Monte Carlo studies have been conducted to examine the properties of fit indices in detecting misfit in covariance structure. However, those studies have primarily considered models in which factor loadings were freely estimated. In GCMs, the factor loadings are usually fixed at a point in time or a point that is a function of time.

Fit indices for the marginal mean structure. SEM-based fit indices may be also used to test the fit of the marginal mean structure. First, the chi-square test statistic is built on the minimized fitting function that

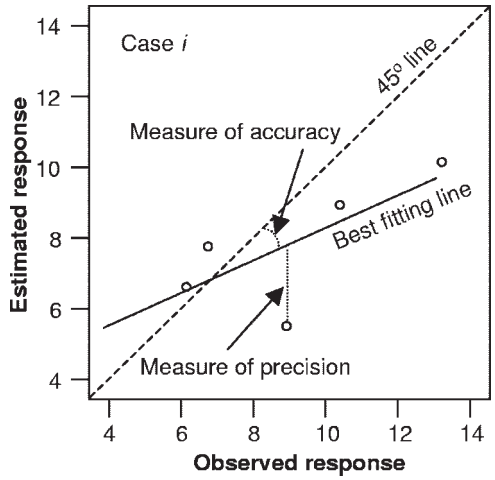

Figure 4. Simplified illustration of concordance correlations (CCCs). The 45° line through the origin indicates perfect fit. The CCC measures (average and conditional CCCs) provide information about the degree of precision (deviation of the individual response from the best fitting line) and the degree of accuracy (angle between the best fitting line and the 45° line) for the fitted model. Hypothetical data are shown for only 1 participant at five measurement occasions.

takes into account the marginal mean structure (the $(\bar{\mathbf{Y}} - \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}))'\hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta})^{-1}(\bar{\mathbf{Y}} - \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}))$ term in Equation 11), so that the chi-square test statistic reflects the fit of the marginal mean structure. Similarly, the practical fit indices based only on the chi-square test statistic for the hypothesized model (e.g., root-mean-square error of approximation [RMSEA] and McDonald's measure of centrality [Mc]), should also reflect the fit of the model to the marginal mean structure.

Second, relative fit indices (e.g., comparative fit index [CFI] and Tucker–Lewis index [TLI]) are defined through the chi-square test statistic, so they should be able to reflect the fit of the model to the marginal mean structure. The major issue for relative fit indices is specification of an acceptable baseline model. Fit indices with an incorrectly specified baseline model have no valid interpretation and may lead to biased inferences. Of importance, the standard baseline model used by nearly all SEM packages (e.g., EQS, Bentler, 1995; LISREL, Jöreskog & Sörbom, 1996; Mplus, L. K. Muthén & Muthén, 1998–2007) is not an appropriate baseline model for commonly used GCMs (Widaman & Thompson, 2003). The specification of an acceptable baseline model for GCMs is discussed in the next section.

Third, residual-based fit indices (e.g., standardized root-mean-square residual [SRMR] and weighted root-mean-square residual [WRMR]) are weighted functions of model residuals. As long as these fit indices take into account the residuals of the marginal means (deviation of the sample means from the model implied means), they should reflect the model fit for the marginal means. So the central issue is whether a residual-based fit index takes into account the residuals in mean structure directly. From examining the formula for SRMR (see Table 2), it is not clear whether it takes into account the residuals in mean structure.

In addition, because the average CCC and pseudo R^2 measure the agreement between the observed responses and model estimated marginal means, both can be used as an index of goodness of fit of marginal mean structure. Compared with the average CCC, one disadvantage of pseudo R^2 is that because it is built on the Pearson correlation, it is not able to detect the deviations of the best fitting lines from the perfect line for the pairs of the observed individual responses and estimated marginal means.

Fit indices for conditional mean structure. The SEM-based fit indices are based on the sample means and the model-implied marginal means instead of individual observed and estimated responses; thus, SEM-based fit indices *cannot* be used to test the fit of the conditional mean structure. In contrast, the MLM-based conditional CCC measures the agreement between the observed and estimated individual responses, so that it can be used to test the fit of conditional mean structure.

Can we construct SEM- and MLM-based fit indices for the three types of longitudinal data? Let us consider SEM-based fit indices first. The ability to obtain SEM fit indices is directly related to the type of longitudinal data being modeled. Raudenbush (2001) noted that a "gold-standard" unrestricted (saturated) model, which is required for most evaluations of model fit in SEM, can be defined only for models that assume homogenous covariance matrices across individuals ($\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$). In covariance structure modeling, the unrestricted model has traditionally been defined having all of the covariance components freely estimated. For GCMs containing both a mean and covariance structure, Browne and Arminger (1995) added the requirement of a homogeneous mean structure across individuals ($\boldsymbol{\mu}_i = \boldsymbol{\mu}$) so that a saturated model may be derived. If the hypothesized model is nested within the saturated model, we can test the fit of the hypothesized model via the likelihood ratio test. Of importance, only with Type I and Type II longitudinal data can homogeneous mean and covariance matrices be constructed. Thus, with these two types of longitudinal data, we can compute overall chi-square tests of model fit and SEM fit indices on the basis of this chi-square test statistic. In addition, although the regular residual-based fit indices do not require a saturated model, they do require a common sample and model implied mean and covariance structure, which are *only* plausible with data from Type I or Type II longitudinal data. Thus, SEM-based fit indices can be only constructed with Type I and Type II longitudinal data.

Now consider the MLM-based fit indices. MLM-based fit indices do *not* require a saturated model. The calculation of

pseudo R^2, average CCC, and conditional CCC involve only observed individual responses and either the model implied marginal means or model implied individual responses, which are available for all of the three types of longitudinal data.

Table 3 summarizes the sources of misfit that theoretically can be detected by SEM- and MLM-based fit indices and for what type of longitudinal data those fit indices can be calculated.

Factors That Affect the Performance of Fit Indices

Many factors have been found to influence the performance of SEM-based fit indices. Here, we focus on three commonly considered factors: model misspecification, distribution, and sample size. The three factors do not exert their effects on the performance of fit indices independently. Hu and Bentler (1998) argued that a good fit index should have a large model misspecification effect accompanied by trivial effects of sample size and distribution. Given space limitations, we briefly summarize past research on the effect of the three factors.

In terms of sensitivity to model misspecification, past research in the context of confirmatory factor analysis models *with only covariance structure* has examined how SEM-based fit indices are related to the misspecification of factor covariances versus misspecification of factor loadings. TLI, CFI, RMSEA, and WRMR were more sensitive to misspecification of the factor loadings than to the misspecifications of factor covariance (Hu & Bentler, 1998; Yu, 2002). In contrast, SRMR showed the opposite effect. Yu (2002) found that CFI, TLI, RMSEA, and SRMR did not overreject trivially misspecified models. WRMR rejected models with trivial misspecification in the factor covariances too frequently under all sample size conditions, and it overrejected models with trivial misspecifications in the factor loadings in the large sample size condition ($N = 1{,}000$). Given that SRMR appeared to show sensitivity to a different type of misspecification than the other fit indices, Hu and Bentler (1998, 1999) recommended a two-index strategy of using SRMR in combination with one of the other fit indices more sensitive to factor loading misspecification (e.g., CFI, RMSEA) to evaluate model fit. However, Fan and Sivo (2005) noted that Hu and Bentler did not quantify the severity of model misspecification and thus confounded type with severity of misspecification in their study. Fan and Sivo showed that, controlling for the severity of misspecification, Hu and Bentler's conclusion that the fit indices were differently sensitive to different types of misspecification no longer held.

In the context of GCMs with five or eight time points, Yu (2002) showed that SEM-based fit indices differ in their power to detect misspecification in GCMs. He found that for the same misspecified model (a linear GCM was fitted to data generated by a quadratic model), the chi-square test statistic, TLI, CFI, and RMSEA all had $> .80$ power to reject the model using suggested cutoff criteria ($T < \chi^2_{\text{critical at } \alpha = .05}$, TLI ≥ 0.95, CFI ≥ 0.95, RMSEA $\leq .07$) when $N \geq 250$. With 5 time points, SRMR had $< .80$ power to reject the misspecified GCM with the cutoff criterion of SRMR $< .07$ when $N \geq 250$. However, SRMR did have adequate power to reject the misspecified GCM with eight time points. The cutoff criteria of WRMR < 1.0 led to reasonable Type I error rates and large power for the GCM with five time points, whereas WRMR tended to overreject the true GCM with eight time points with this cutoff criterion.

Leite and Stapleton (2006) investigated the sensitivity of RMSEA and SRMR to misspecification of the functional form of growth (fitting a linear functional form to the data generated from nonlinear and piecewise GCMs). They manipulated the severity of misspecification (slight, moderate, and strong) defined on the basis of power calculations (Saris & Satorra, 1993; Satorra & Saris, 1985) for the alternative models. They found that RMSEA was very sensitive to misspecification in functional form using a cutoff criterion of RMSEA $\leq .06$ (average power = .89, .97, and 1.00 for slight, moderate, and strong misspecification, respectively), whereas SRMR had unacceptably low power to reject models with misspecified mean and covariance structures with a cutoff criterion of SRMR $\leq .08$ (average power = .12, .41, and .36 for slight, moderate, and strong misspecification, respectively). As a comparison, they tested the same linear models when the mean structure was specified as saturated and would not contribute to misfit. Under these conditions, SRMR had higher power to reject the misspecification in covariance structure (average power = .31, .71, and .83 for slight, moderate, and strong misspecification, respectively). The average power for RMSEA under the different levels of severity of misspecification did not change substantially when the mean structure was specified as saturated. Thus, it

Table 3
Fit Indices to Detect the Misfit in Marginal Mean, Covariance, and Conditional Mean Structure

Fit indices	Sources of misfit	Type of longitudinal data
SEM-based fit indices[a]	Marginal mean structure and covariance structure	Types I and II
Average CCC, Pseudo R^2	Marginal mean structure	Types I, II, and III
Conditional CCC	Conditional mean structure	Types I, II, and III

Note. SEM = structural equation modeling; CCC = concordance correlation.
[a] Residual-based fit indices (e.g., standardized root-mean-square residual) appear to be less sensitive to the marginal mean structure than the likelihood-ratio-based fit indices.

appears that SRMR is less sensitive to misspecification in mean structure than to the misspecification in covariance structure.

As described earlier, GCMs often assume multivariate normality of the longitudinal responses and random effects. Given multivariate normality, the ML-based chi-square test statistic (T_{ML}) asymptotically follows a central or noncentral chi-square distribution. In practice, the multivariate normality assumption is often violated. Under these conditions, T_{ML} does *not* follow a chi-square distribution; therefore, inferences based on T_{ML} may have an inflated Type I error rate (Curran, West, & Finch, 1996; Satorra, 1992; Yu, 2002).

To improve the approximation of the chi-square test statistic to a chi-square distribution under conditions of nonnormal data, Satorra and Bentler (1988) developed a rescaled chi-square test statistic (T_{SB}). T_{SB} corrects T_{ML} by a constant k, $T_{SB} = T_{ML}/k$ (Bentler & Yuan, 1999). T_{SB} asymptotically follows a chi-square distribution if the observed variables have homogeneous marginal skewness and kurtosis. With heterogeneous marginal skewness and kurtosis, the asymptotic distribution of T_{SB} is unknown (Bentler & Yuan, 1999). Instead it will approach a variate with expected value equal to *df*. T_{SB} rescales the chi-square test statistic by correcting only for the mean of the distribution (Bentler & Yuan, 1999; Yuan, 2005; Yuan & Bentler, 1998). T_{SB} behaved extremely well relative to T_{ML} under nonnormal distributions, except that it tends to overreject correct models slightly when sample sizes are small (about $N = 125$; Bentler & Yuan, 1999; Curran et al., 1996; Yu, 2002; Yuan & Bentler, 1998).

There are few studies on fit indices based on the chi-square test statistic other than T_{ML}. In the context of covariance structure only models, Yu (2002) showed that T_{SB}-based CFI, TLI, and RMSEA led to a lower Type I error rate than T_{ML}-based CFI, TLI, and RMSEA when $N \leq 250$. T_{ML}-based CFI, TLI, and RMSEA had higher power than T_{SB}-based CFI, TLI, and RMSEA under almost all the sample size conditions, but the higher power associated with ML-based fit indices was suspect because T_{ML} was inflated with increasing nonnormality for misspecified models.

Strictly speaking, there is no fit index that is completely unaffected by sample size. Sample size exerts its potential effect on model fit indices in two ways. First, sample size may enter the calculation of a fit index directly so that the value of a fit index increases as sample size increases, which will lead to an inflated Type I error rate (Bollen, 1989; Sun, 2005). For example, $T_{ML} = (N - 1)F_{ML}$ tends to be inflated by sample size when the model is misspecified (Bollen, 1989). As a result, even a trivially misspecified model could be rejected by this test under some commonly seen conditions. Second, an increased Type I or Type II error rate may be introduced at small sample sizes because the asymptotic distribution is not well approximated when sample size is small (Bollen, 1989; Sun, 2005). Some fit indexes are much less sensitive to sample size than others. For example, population-based fit indices (e.g., RMSEA) are expected to be less affected by sample size than sample-based fit indices, and relative fit indices (e.g., CFI and TLI) are expected to be less affected by sample size than absolute fit indices. Fan and Sivo (2005) found that sample size had more effect on fit indices under slight misspecification than under moderate misspecification.

The factors of misspecification, distribution, and sample size may also affect MLM-based fit indices. However, there is a lack of systematic research examining the effect of those factors on MLM-based fit indices in the context of GCMs. Many questions remain to be answered. For instance, do MLM fit indices decrease monotonically as model misspecification in mean structure increases? Which fit index is most sensitive to misspecification in marginal mean structure, average CCC, Pseudo R^2, or the SEM-based fit indices? How sensitive is the conditional CCC to misspecification in the conditional mean structure? How is the sensitivity of those fit indices moderated, if at all, by other factors such as sample size or distribution?

It is important to note that both SEM- and MLM-based fit indices will be affected by still other factors when evaluating mean structure. For example, they will be affected by the reliability of the repeated measures. If other conditions are constant, higher reliability of repeated measures will result in higher values of the fit indices for mean structure. In addition, fit indices may be also affected by the extent to which the repeated measures are predicted by the terms included in the Level-1 model. If reliability of measurement is kept constant, cases in which terms related to time (e.g., *time*, $time^2$) can predict a larger rather than smaller proportion of variance in the outcome variable in the population will result in higher values on the fit indices. Consider a GCM with time as the only time-varying predictor in the model. The within-individual covariance matrix (the Level-1 residual covariance matrix) is affected by two factors: (a) the measurement error due to the unreliability of measurement and (b) the variances and covariances among residuals that are accounted for by other time-varying covariates that are not included in the model. Given that the between-individuals covariance matrix is constant, the larger the values in the Level-1 residual covariance matrix, the lower the proportion of variance in the outcome variables that can be accounted for by time, and the smaller the fit indices for marginal means will be. Similarly, the larger the values in the Level-1 residual covariance matrix, the smaller the ratio of between to within-individual variability, the greater the discrepancy between the observed and estimated individual responses, and thus the smaller the conditional CCC will be. As a result, specifying an optimal

functional form[6] in different cases may lead to different values on the fit indices for mean structure.

Issues and Challenges in Evaluating Model Fit for GCMs

Given that GCMs involve both mean and covariance structures and that longitudinal data have different types, several complex issues arise in evaluating model fit for GCMs.

Specifying Correct Baseline Model(s) for Growth Curve Models for Relative Fit Indices

Specifying a correct baseline model is critical for relative fit indices. Fit indices with an incorrectly specified baseline model have no valid interpretation and may lead to biased inferences. Widaman and Thompson (2003) defined the necessary characteristics for an acceptable baseline model. Of critical importance, a baseline model must be *nested* within the hypothesized model under consideration for a set of data. In addition, it must (a) estimate as few parameters as are reasonable for the data, and (b) reproduce a nonzero variance and mean (if included in the analysis) for each manifest variable.

The standard baseline model used by nearly all SEM packages (e.g., LISREL, Mplus) is an independence model in which the covariances among the manifest variables are set to zero, but variances are unrestricted, and means are unrestricted if mean structure is included in the model[7] (Bentler & Bonett, 1980; Widaman & Thompson, 2003). Unfortunately, this standard baseline model is *inappropriate* for growth curve models. Using the linear GCM in Figure 1 as an example, the linear GCM has six parameters to be estimated: means and variances for the intercept and slope, the covariance between intercept and slope, and a constant residual variance for the manifest variables. The traditional baseline model for the linear GCM has eight free parameters: four means and four variances for the manifest variables. It is obvious that the standard baseline model is *not* nested within the linear GCM. Widaman and Thompson (2003) specified an acceptable baseline model[8] that may be used for most of the commonly used GCMs (i.e., all of the polynomial models and linear piecewise models). This baseline model is based on an intercept-only growth model: Only the mean of the intercept and the residual variances for the manifest variables are free parameters. Note that if there is an equality constraint on the residual variances in the hypothesized model (as is typical in the MLM approach), the baseline model must also reflect that constraint. The variance of the latent variable intercept is fixed to be 0 in the baseline model. Figure 5 shows an acceptable baseline model for the above linear GCM, which has two parameters to be estimated: a mean for the intercept and a single constant residual variance for the manifest variables.

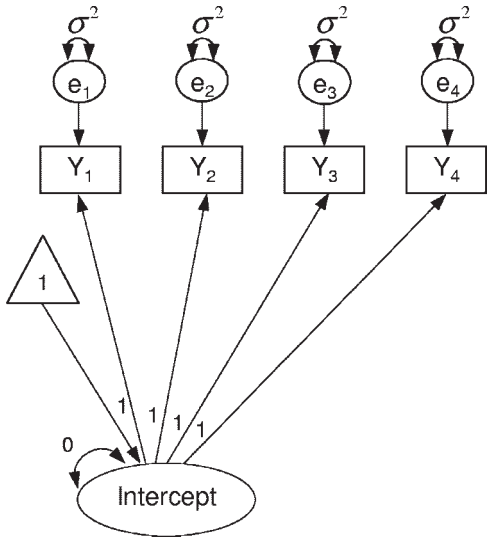

Figure 5. An acceptable baseline model for the linear latent growth curve model in Figure 1. There is only an intercept factor in the model. The free parameters in the model are the mean of intercept and the unique variances of Y_1 to Y_4, which are constrained to be equal corresponding to the hypothesized model in Figure 1. In the model portrayed, the triangle indicates that the mean of the intercept is estimated.

Correcting the relative fit indices requires three steps:

1. Estimate the hypothesized model using any software capable of estimating GCMs (e.g., EQS, LISREL, MPlus) to obtain T_h and df_h for the model but ignore the relative fit index values output by the program;
2. Estimate the correct baseline model to obtain T_b and df_b;
3. Use the values of T_h, df_h, T_b, and df_b to calculate the value of the fit index, for example, TLI =

[6] If we have a limited number of repeated measures for each individual, many forms of growth cannot be estimated. For example, if one has only three repeated measures, one cannot estimate quadratic growth. Such limitations may lead to the adoption of a simple functional form that does not adequately represent the growth process. However, that functional form might be the optimal form that one can specify given the data.

[7] In the standard baseline model used in AMOS (SPSS, 2008), the means of the manifest variables are constrained to equal 0 if mean structure is included in the model. This difference does not affect our conclusion that standard SEM software does not use acceptable baseline models unless unique variances in the test model are all freely estimated.

[8] Sobel and Bohrnstedt (1985) have noted that more than one acceptable baseline model may exist for a specific hypothesized model.

$[(T_b/df_b) - (T_h/df_h)]/[(T_b/df_b) - 1]$. We provide macros in SAS (SAS Institute, 2001) and SPSS (SPSS, 2005) to calculate relative fit indices, including the TLI, normed fit index (NFI), CFI, and parsimony adjusted CFI. These are available in the supplemental materials.

Differentiating Misspecification in the Marginal Mean and Covariance Structures

As described previously, SEM-based fit indices can be used to detect misspecification in both the marginal mean and covariance structures. Previous research on SEM-based fit indices for GCMs only considered joint misspecification involving both the mean and covariance structures. For example, Yu (2002) created misspecified models by dropping the quadratic effect from a quadratic population model. The population model in Leite and Stapleton (2006) was nonlinear or piecewise, and the misspecified model was linear. In both studies, the misspecification in marginal mean and covariance structure were confounded. As a result, the source of misfit is unclear, so no conclusions can be reached about how fit indices reflect misspecifications only in the mean or covariance structure.

It is possible that some fit indices might have differential sensitivity to misspecification only in the mean structure relative to only the covariance structure. If so, one could combine information provided by different fit indices to make an inference regarding the source of model misspecification. For example, on the basis of the results of Leite and Stapelton (2006), the RMSEA might be much more sensitive to misspecification in mean structure than is the SRMR. If one obtained a RMSEA for the tested model that suggested rejection, whereas the SRMR suggested acceptance, then the misspecification is very likely to come from the mean structure.

An alternative potential way to differentiate the misspecification in marginal mean and covariance structure is to specify the covariance structure to be saturated while examining the fit of the marginal mean structure. Comparing different competing functional forms (marginal mean structures) when the role of misspecification in the covariance structure is reduced may give a clearer picture of misspecification in the mean structure. Similarly, the model for the marginal means could be saturated while examining the fit of the covariance structures. Once again, comparing different plausible covariance structures when the marginal means are freely estimated may give a clearer picture of misspecification in the covariance structure. The goal is to estimate a model in which only misspecification in one structure would contribute to the misfit of the model. However, only preliminary work by Leite and Stapleton (2006) has examined the performance of fit indices to detect the misspecification in the covariance structure when the mean structure is specified as saturated.

Difficulties in Determining Cutoff Criteria for Models Involving Mean Structures

Indices based on both MLM and SEM can be used to evaluate the fit of the marginal mean structures. For SEM-based fit indices, it has been standard practice to evaluate model fit indices using rule of thumb cutoff criteria (e.g., CFI = .95, RMSEA = .06, SRMR = .08; see Hu & Bentler, 1998, 1999). However, these rule of thumb cutoff criteria were developed on the basis of practical experience and simulation studies using covariance structure models *without* a mean structure. Yu (2002) found that only some of the criteria were appropriate for GCMs with misspecifications in both mean and covariance structures. However, the conditions investigated by Yu confounded misspecification in the mean and covariance structures. If we wish to focus on detecting misspecification only in the marginal mean structure, then it is likely that such rule of thumb guidelines do not apply. Consider trying to develop potential cutoff criteria for regression analysis or trend analysis in analysis of variance, two analysis approaches in which the marginal mean structure is of primary importance. There is little consensus among researchers regarding what would constitute an acceptable level of fit.

For MLM-based fit indices, no such cutoff criteria have been proposed. The existing applications of the MLM-based fit indices often use fit indices to compare competing models instead of to indicate the global fit directly. For example, to model longitudinal data on the number of epileptic seizures over time (Thall & Vail, 1990), Vonesh et al. (1996) used the average CCC to compare competing longitudinal models with different marginal functional forms. They then used the selected marginal functional form to choose an appropriate covariance structure. Singer and Willet (2003) proposed that researchers can evaluate the contribution of an added predictor to a model by using the increase in pseudo R^2 associated with the predictor. In addition, the MLM fit indices presented in the current article do not adjust for model complexity, thus they will always increase as model complexity increases.

The Effect of Missing Data on Fit Indices

Type II data in which some data are MAR create complications in both the estimation and evaluation of model fit in GCMs. With Type II data, missing data techniques—such as FIML and multiple imputation (MI)—produce unbiased estimates of the mean growth trend after conditioning on the observed data (see Schafer & Graham, 2002). However, as noted by Enders (2001), when FIML is used for Type II longitudinal data, the chi-square test statistic *cannot* be calculated using the general form, $(N - 1)F_{FIML}$, because there is no

single N that is applicable to the entire sample. Researchers can calculate the fit indices by following a two-step procedure: (a) estimate the saturated model and the hypothesized model using FIML, and (b) calculate the chi-square test statistics using the formula $-2(\ln L_{FIML(hypothesized)} - \ln L_{FIML(saturated)})$. Then, the fit indices based on the chi-square test statistics can be calculated by hand using standard formulas (see Table 3).

With MI, a Monte-Carlo-based technique is used to impute plausible values for the missing data in multiple data sets. Each data set can be analyzed separately using standard methods developed for complete data. All of the standard estimation methods and model fit indices applied to complete data can be used in each data set and then recombined. However, in some cases, complexities may arise in the specification of the proper N and df. In addition, there may be convergence problems associated with MI that become far more likely with some patterns of missing data (e.g., the multiple cohort design).

Although we can still calculate SEM-based fit indices when data are MAR, such data raise problems for the performance of model fit indices. Davey, Savla, and Luo (2005) found that missing data greatly reduced the ability of SEM-based fit indices (NFI, CFI, RMSEA, and Mc) to detect misspecified *covariance structure* models. Using FIML estimation, they found that when the data were missing completely at random (MCAR), $F_{\min}$ decreased linearly. In contrast, when the data were MAR, $F_{\min}$ decreased nonlinearly. To date, there is no good solution to the problem of evaluating model fit when data are missing. There is also little research on the topic, and none in the context of GCMs. Given the linear relationship between $F_{\min}$ and the proportion of missing data in covariance structure models, it may be possible to correct fit indices for the proportion of missing data when data are MCAR. Alternatively, relative rather than the absolute values of fit indices can be used to choose the best of a set of competing models. Tests indicating that missing data are consistent with an MCAR structure are available (Chen & Little, 1999; Park & Lee, 1997). However, MCAR data may be rare in practice. To date, no study has examined the effect of missing data on the MLM-based fit indices.

Evaluating Model Fit for GCMs With Type III Data

As discussed previously, SEM-based fit indices *cannot* be calculated for Type III longitudinal data. One is not able to evaluate how the model estimated marginal mean and the covariance structures match the sample mean vector and covariance structure. However, one can still evaluate the fit of the marginal means using the pseudo R^2 and average CCC as well as fit of the conditional means using the conditional CCC. The challenge is how those fit indices will be affected by the imbalances.

An alternative (and complementary) way to examine the fit of growth models is to adapt methods for the examination of residuals from multiple regression analysis (Cohen et al., 2003; Weisberg, 2005). Fitzmaurice et al. (2004) and Weiss (2005) have both noted that the analysis of longitudinal data is not complete without a close examination of the residuals. Residual diagnoses can not only help check the adequacy of the mean and covariance structures but also help detect outliers and violations of assumptions—for example, heterogeneity of variance, serial dependency (autocorrelation) among Level-1 residuals, and nonnormality of the random effects. Given space limitations, we cannot consider residual diagnosis in detail in this article. Interested readers should refer to Weiss (2005, pp. 327–341) and Fitzmaurice et al. (2004, pp. 237–253).

Conclusions and Recommendations

This article has provided an overview of fit indices from both the SEM and MLM frameworks that can be used to evaluate model fit for GCMs and the challenges in their use. Although MLM and SEM represent different approaches to GCMs, in most applications they share identical model-implied mean and covariance matrices, statistical assumptions, and estimation methods (Mehta & West, 2000). However, the measurement of goodness of fit has historically differed between the approaches. SEM has emphasized measuring the match between sample and model implied mean responses and the match between the sample and model implied covariance matrices using the chi-square test statistic and other practical fit indices. MLM has emphasized measures of the agreement between the estimated mean or estimated individual responses with the observed individual responses using different correlational measures.

The first issue in choosing and interpreting model fit statistics is the type of longitudinal data structure. Following Raudenbush (2001), three types of data structures in longitudinal data were distinguished: (a) Type I balanced design with complete data, (b) Type II balanced design with MAR data, and (c) Type III unbalanced design (see Table 1). All SEM and MLM fit indices can be calculated for Type I longitudinal data. Although all SEM and MLM fit indices can be calculated for Type II longitudinal data, missing data have important effects on their interpretation. Missing data reduce the power of SEM-based fit indices to differentiate between correctly and incorrectly specified models. The fit indices show attenuated ability to detect sources of misfit, and traditional guidelines for adequate fit may be misleading. Type III longitudinal data severely restrict the set of available fit indices. None of the SEM-based fit indices can be calculated. All of the MLM-based fit indices—including the pseudo R^2, the average CCC, and the conditional CCC—are potentially available.

The second issue is that growth curve models involve multiple sources of misspecification. We identified four sources of misspecification in GCMs: within-individual covariance structure, between-individuals covariance structure, marginal mean structure, and conditional mean structure. Only SEM-based fit indices directly reflect misspecification in the covariance structures. Only the MLM-based conditional CCC reflects the misspecification in the conditional mean structure. Both SEM- and MLM-based fit indices reflect misspecification in the marginal mean structure. The fit of the mean structures, especially the fit of the conditional mean structure, has received less emphasis than the fit of the covariance structures. The fit of the conditional mean structure in addition to the fit of marginal mean structure needs to be considered to find the appropriate functional form for a GCM. In addition, the fit of the total covariance structure cannot guarantee that the between-individuals and within-individual variability in the outcome has been correctly captured.

It is very challenging to differentiate the different sources of misspecification. One strategy may be to utilize combinations of fit indices from both the MLM and SEM approaches that reflect different sources of misfit. This strategy could provide a fuller picture of the adequacy of fit of GCMs. A second possible strategy is to specify either the mean or covariance structure to be saturated so that the influence of that structure is minimized and different specifications in the other structure can be compared. The extent to which each of these strategies can detect various forms of misspecification in the mean and covariance structures awaits future research.

Given the caveat that virtually all of the existing work has been conducted using *balanced designs with complete data* (*Type I data*), the RMSEA, CFI, and TLI among the SEM-based fit indices have shown good potential performance in evaluating the fit of GCMs. These fit indices (a) can be applied to models with both mean and covariance structures, (b) are sensitive to model misspecification without overrejection of true or trivially misspecified models, and (c) are minimally affected by sample size. The RMSEA is affected by nonnormality and high correlations among the measured variables. The proper use of the CFI and TLI requires that a correct baseline model be manually estimated (see the supplemental materials). In the context of GCMs, researchers should be cautious about using the WRMR and SRMR, considering that the WRMR may lead to overrejection of trivially misspecified models, and the SRMR generally had low power to detect misspecification in mean structure.

When data meet the assumption of multivariate normality (or more precisely, asymptotic robustness), the ML chi-square test provides the optimal test of the exact fit of the overall model to the data for moderate to large sample sizes. The Satorra–Bentler test statistic has superior performance to the ML chi-square when there is nonnormality in the data for moderate to large sample sizes. The Satorra–Bentler test statistic is available in nearly all SEM programs. However, the studies on fit indices in the context of GCMs have only considered the performance of fit statistics using standard ML estimation, precluding definitive statements about their performance when alternative robust estimators (e.g., T_{SB}) are used.

We presented three MLM-based fit indices in this article. Compared with SEM-based fit indices, the MLM-based fit indices can be used with unbalanced designs, and the conditional CCC can be used to detect misspecification in conditional mean structure. However, the properties of MLM-based fit indices have not been extensively studied in the context of GCMs. These fit indices have good theoretical properties, but it is an open question whether their actual performance in practice with GCMs will fully match their theoretical properties.

Given the multiple potential sources of misfit and the three types of longitudinal data structures, evaluating goodness of fit is much more difficult for GCMs than for covariance structure models. Researchers have typically evaluated the fit of covariance structure models by comparing SEM-based fit indices with conventional cutoff criteria. For GCMs, fit indices will also be affected by aspects of the mean structure. The fit of mean structure is not only affected by the misspecification in mean structure but also by the reliability of measurement, the range of the time variable, and the extent to which the outcome variable is predicted by the time variable and any other time-varying predictor(s) of interest. It may be impossible to establish cutoff criteria that are tenable for models with varying reliability and extent to which the outcome variable is predicted by time. Rather than attempting to make appeals to conventional cutoff criteria associated with each model fit index, it is important to use fit indices that best reflect the central questions of the researchers. For example, if the central question of a researcher is to identify the optimal functional form relating time to the outcome variable of interest, a good strategy may be to examine combinations of fit indices that are differentially sensitive to the marginal and conditional mean structures. Through the use of such strategies, researchers will have greater ability to identify the aspects of the mean and covariance structures that are the source of misfit in their models.

References

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238–246.

Bentler, P. M. (1995). *EQS: Structural equations program manual.* Encino, CA: Multivariate Software.

Bentler, P. M., & Bonett, D. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588–602.

Bentler, P. M., & Yuan, K.-H. (1999). Structural equation modeling with small samples: Test statistics. *Multivariate Behavioral Research, 34,* 181–197.

Biesanz, J. C., West, S. G., & Kwok, O.-M. (2003). Personality over time: Methodological approaches to the study of short-term and long-term development and change. *Journal of Personality, 71,* 905–941.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective.* New York: Wiley.

Browne, M. W., & Arminger, G. (1995). Specification and estimation of mean and covariance structure models. In G. A. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 185–249). New York: Plenum Press.

Browne, M. W., & Du Toit, S. H. C. (1991). Models for learning data. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change* (pp. 47–68). Washington, DC: American Psychological Association.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101,* 147–158.

Chen, H. Y., & Little, R. (1999). A test of missing completely at random for generalised estimating equations with missing data. *Biometrika, 86,* 1–13.

Chou, C., Bentler, P. M., & Pentz, M. A. (1998). Comparison of two statistical approaches to study growth curves: The multilevel model and the latent curve analysis. *Structural Equation Modeling, 5,* 247–266.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Curran, P. J., & Peterman, M. (2005, October). *A curious discrepancy between multilevel and structural equation growth curve models with time-varying covariates.* Paper presented at annual meeting of the Society for Experimental Psychology, Lake Tahoe, CA.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1,* 16–29.

Davey, A., Savla, J., & Luo, Z. (2005). Issues in evaluating model fit with missing data. *Structural Equation Modeling, 12,* 578–597.

Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling, 8,* 128–141.

Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indices to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling, 12,* 343–367.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis.* New York: Wiley.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Arnold.

Graham, J. W., Taylor, B. J., Cumsille, P. E., & Olchowski, A. E. (2006). Planned missing data designs in psychological research. *Psychological Methods, 11,* 323–343.

Hoyle, R. H. (1995). *Structural equation modeling.* Thousand Oaks, CA: Sage.

Hu, L.-T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3,* 424–453.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

James, L. R., Mulaik, S. A., & Brett, J. (1982). *Causal analysis: Models, assumptions, and data.* Beverly Hills, CA: Sage.

Jöreskog, K. G., & Sörbom, D. (1981). *LISREL VI: Analysis of linear structural relationship by maximum likelihood and least squares method.* Chicago: National Educational Resources.

Jöreskog, K. G., & Sörbom, D. (1984). *LISREL VI user's guide.* Mooresville, IN: Scientific Software.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8 user's reference guide.* Chicago: Scientific Software.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics, 38,* 963–974.

Leite, W. L., & Stapleton, L. M. (2006). Sensitivity of fit indices to detect misspecifications of growth shape in latent growth modeling. Retrieved from http://plaza.ufl.edu/leitewl/Presentation_AERA2006_misspecification_of_growth.pdf

Lin, L. A. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics, 45,* 255–268.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

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.

Marsh, H. W., Hau, K. T., & Wen, Z. L. (2004). In search of golden rules: Comment on hypothesis testing approaches to setting cutoff values for fit indexes and dangers in overgeneralising Hu and Bentler (1999) findings. *Structural Equation Modeling, 11,* 320–341.

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification, 6,* 97–103.

Mehta, P., & West, S. G. (2000). Putting the individual back in individual growth curves. *Psychological Methods, 5,* 23–43.

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

Muthén, B. O., & Curran, P. J. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods, 2,* 371–402.

Muthén, B. O., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52,* 431–462.

Muthén, L. K., & Muthén, B. O. (1998–2007). *Mplus user's guide* (5th ed.). Los Angeles: Author.

Park, T., & Lee, S.-Y. (1997). A test of missing completely at random for longitudinal data with missing observations. *Statistics in Medicine, 16,* 1859–1871.

Rao, C. R. (1965). The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika, 52,* 447–468.

Raudenbush, S. W. (2001). Toward a coherent framework for comparing trajectories of individual change. In L. Collins & A. Sayer (Eds.), *New methods for the analysis of change* (pp. 33–64). Washington, DC: American Psychological Association.

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

Saris, W. E., & Satorra, A. (1993). Power evaluations in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 181–204). Newbury Park, CA: Sage.

SAS Institute. (2001). *Statistical analysis system.* Available at http:/www.sas.com

Satorra, A. (1992). Asymptotic robust inferences in the analysis of mean and covariance structures. *Sociological Methodology, 22,* 249–278.

Satorra, A., & Bentler, P. M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. In *ASA 1988 Proceedings of the Business and Economic Statistics Section* (pp. 308–313). Alexandria, VA: American Statistical Association.

Satorra, A., & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika, 50,* 83–90.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York: Oxford.

Sobel, M. E., & Bohrnstedt, G. W. (1985). Use of null models in evaluating fit of covariance structure models. *Sociological Methodology, 15,* 152–178.

SPSS. (2005). *SPSS Base 14.0 user*'s *guide.* Chicago: Author.

SPSS. (2008). *Amos 16.0 user's guide*. Chicago: Author.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically based tests for the number of common factors.* Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika, 50,* 253–263.

Sun, J. (2005). Assessing goodness of fit in confirmatory factor analysis. *Measurement and Evaluation in Counseling and Development, 37,* 240–256.

Thall, P. F., & Vail, S. C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics, 46,* 657–671.

Tucker, L. R. (1958). Determination of parameters of a functional relation by factor analysis. *Psychometrika, 23,* 19–23.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1–10.

Verbeke, G., & Lesaffre, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis, 23,* 541–556.

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

Vonesh, E. F., & Chinchilli, V. M. (1997). *Linear and non-linear models for the analysis of repeated measures.* Basel, Switzerland: Marcel Dekker.

Vonesh, E. F., Chinchilli, V. M., & Pu, K. (1996). Goodness-of-fit in generalized nonlinear mixed-effects models. *Biometrics, 52,* 572–587.

Ware, J. H. (1985). Linear models for the analysis of longitudinal studies. *American Statistician, 39,* 95–101.

Weisberg, S. (2005). *Applied linear regression* (3rd ed.). Hoboken, NJ: Wiley.

Weiss, R. E. (2005). *Modeling longitudinal data.* New York: Springer.

Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods, 8,* 16–37.

Willett, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin, 116,* 363–381.

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes.* Unpublished doctoral dissertation, University of California, Los Angeles.

Yuan, K.-H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research, 40,* 115–148.

Yuan, K.-H., & Bentler, P. M. (1998). Normal theory based test statistics in structural equation modeling. *British Journal of Mathematical and Statistical Psychology, 51,* 289–309.

Yuan, K.-H., & Bentler, P. M. (2004). On chi-square difference and *z* tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological Measurement, 64,* 737–757.

Received June 5, 2007
Revision received December 22, 2008
Accepted January 6, 2009 ■