Correcting Model Fit Criteria for Small Sample Latent Growth Models With Incomplete Data

Educational and Psychological
Measurement

2017, Vol. 77(6) 990–1018

© The Author(s) 2016

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164416661824

journals.sagepub.com/home/epm

Daniel McNeish^{1,2} and Jeffrey R. Harring¹

Abstract

To date, small sample problems with latent growth models (LGMs) have not received the amount of attention in the literature as related mixed-effect models (MEMs). Although many models can be interchangeably framed as a LGM or a MEM, LGMs uniquely provide criteria to assess global data–model fit. However, previous studies have demonstrated poor small sample performance of these global data–model fit criteria and three post hoc small sample corrections have been proposed and shown to perform well with complete data. However, these corrections use sample size in their computation—whose value is unclear when missing data are accommodated with full information maximum likelihood, as is common with LGMs. A simulation is provided to demonstrate the inadequacy of these small sample corrections in the near ubiquitous situation in growth modeling where data are incomplete. Then, a missing data correction for the small sample correction equations is proposed and shown through a simulation study to perform well in various conditions found in practice. An applied developmental psychology example is then provided to demonstrate how disregarding missing data in small sample correction equations can greatly affect assessment of global data–model fit.

Keywords

latent growth model, small sample, missing data, dropout, full information maximum likelihood, FIML, correction

¹University of Maryland, College Park, MD, USA

²Utrecht University, Utrecht, Netherlands

Corresponding Author:

Daniel McNeish, Department Methodology and Statistics, Utrecht University, PO Box 80140, 3508 TC Utrecht, Netherlands.

Email: d.m.n.mcneish@uu.nl

When using growth models as is common in many disciplines, researchers often account for the covariance between repeated measures with either mixed-effect models (MEMs), or latent growth models (LGMs). The properties of MEMs have recently been studied under small sample size conditions (see, e.g., Bell, Morgan, Schoeneberger, Kromrey, & Ferron, 2014; W. J. Browne & Draper, 2006; Maas & Hox, 2005). Unlike MEMs, no studies have been extended that explicitly focus on small sample issues with LGMs despite their popularity in behavioral sciences (Bollen & Curran, 2006).

Although some parameterizations of MEMs and LGMs can lead to many similarities between the different types of models (Curran, 2003), LGMs have the added benefit that global model fit can be assessed (Chou, Bentler, & Pentz, 1998; Wu, West, & Taylor, 2009). As will be discussed in more detail in subsequent sections, global model fit statistics and indices, which are commonly used with LGMs, are problematic with small sample sizes and vastly overreject models that fit the data well in actuality (e.g., Bentler & Yuan 1999; Kenny & McCoach, 2003; Nevitt & Hancock, 2004; Yuan & Bentler, 1999). To combat this issue, several post hoc small sample corrections have been developed (Bartlett, 1950; Swain, 1975; Yuan, 2005) and have been shown to perform well with structural equation models broadly construed (of which LGMs are a special case) with complete data (Fouladi, 2000; Herzog & Boomsma, 2009; Nevitt & Hancock, 2004). However, each of these post hoc corrections includes sample size in their formulations. With LGMs and structural equation models generally, missing data are often accommodated with full information maximum likelihood (FIML) which does not impute values for missing data but rather makes optimal use of the values that were directly observed. For instance, a sample of 100 participants that treats missing values with FIML may not really contain a full “100 participants’ worth” of information.

This is meaningful when considering the post hoc small sample corrections with missing data because, by using the full sample size in the computational equations for the small sample corrections, the corrections assume that more information is present than there is in reality. As a result, the post hoc corrections no longer provide a viable solution with missing data because the correction is not large enough. For instance, consider a scenario with 100 participants and five repeated measures but imagine that the data are collected in accordance with a planned missingness design such that participants’ responses are not collected for two of the five time points (though not necessarily the same two time points for each individual). These data would contain roughly “60 participants’ worth” of information (depending on the growth trajectory) when FIML is applied; however, the post hoc small sample corrections will use $n = 100$ but this is far too large when considering the true impact of the missing values. When missing data are accounted for with FIML, 100 complete observations contain much more information than 100 observations each missing two time points. Yet current small sample corrections do not distinguish these two scenarios and treat them identically despite the fact that the amount of information provided by the data can be quite different.

After discussing the aforementioned concepts in detail, this article will provide a simulation study to investigate the extent to which missing data affect the viability of post hoc small sample adjustments. Subsequently, a missing data correction for sample size within the post hoc corrections for small samples in the vein of the work by Rubin and Schenker (1986) for multiple imputation will be suggested and its performance explored. A developmental psychology example is then provided to show the impact of (a) ignoring small sample sizes in data-model fit assessment with LGMs and (b) ignoring missing data in small sample corrections to data-model fit statistics and indices.

Brief Introduction to Latent Growth Models

The general linear LGM with k time-invariant covariates can be written as a confirmatory factor analysis (CFA) model with an imposed mean structure such that

$$\begin{aligned} Y_{ij} &= \eta_{0i} + \eta_{1i}t_{ij} + \varepsilon_{ij} \\ \eta_{0i} &= \alpha_{\beta 0} + \gamma_{01}X_{1i} + \cdots + \gamma_{0k}X_{ki} + \zeta_{0i} \\ \eta_{1i} &= \alpha_{\beta 1} + \gamma_{11}X_{1i} + \cdots + \gamma_{1k}X_{ki} + \zeta_{1i}, \end{aligned} \quad (1)$$

where Y_{ij} is the response for the i th individual at the j th time, η_{0i} is the latent intercept for the i th individual, η_{1i} is the latent slope for the i th individual, $\zeta_i = (\zeta_{0i}, \zeta_{1i})^T$ is a vector of factor scores (random effects) for the i th individual, t_{ij} is the j th time point for the i th individual, and ε_{ij} is the residual for the i th individual at the j th time. In matrix notation, Equation (1) becomes

$$\mathbf{Y}_i = \mathbf{\Lambda}_i \boldsymbol{\eta}_i + \mathbf{E}_i \quad (2)$$

and

$$\boldsymbol{\eta}_i = \boldsymbol{\alpha} + \mathbf{\Gamma} \mathbf{X}_i + \boldsymbol{\zeta}_i. \quad (3)$$

The model-implied mean and covariance structures of the repeated measures are thus,

$$\boldsymbol{\mu}_i = \mathbf{\Lambda}_i (\boldsymbol{\alpha} + \mathbf{\Gamma} \boldsymbol{\kappa}) \quad (4)$$

and

$$\boldsymbol{\Sigma}_i = \mathbf{\Lambda}_i (\mathbf{\Gamma} \boldsymbol{\Phi} \mathbf{\Gamma}^T + \boldsymbol{\Psi}) \mathbf{\Lambda}_i^T + \boldsymbol{\Theta}_i, \quad (5)$$

where $\boldsymbol{\mu}_i$ is a vector of model-implied means of the outcome variables, $\mathbf{\Lambda}_i$ is a matrix of loadings that can, but are not always, prespecified to fit a specific type of growth trajectory, $\boldsymbol{\alpha}$ is a vector of latent factor means, $\boldsymbol{\Sigma}_i$ is the model-implied covariance matrix of the outcome variables, $\boldsymbol{\Psi}$ is the covariance matrix of the random effects, $\boldsymbol{\Theta}_i$ is a matrix of residual variances and covariances among the repeated

measures, Γ is a matrix of coefficients for the predicted effect of time-invariant covariates on the latent growth trajectory factors, κ is a vector of covariate means, and Φ is a covariance matrix of the covariates X_{ki} (Biesanz, Deeb-Sossa, Papadakis, Bollen, & Curran, 2004; Curran, 2003). For readers familiar with HLM notation in MEMs, Equation (1) looks similar to the HLM model specification. However, LGMs possess an advantage over MEMs in that they output global data–model fit criteria (Wu et al., 2009).

Fit Criteria for LGMs and Small Samples

Data–Model Fit Criteria

While fit in MEMs is typically assessed through inferential tests on individual model parameters (Wu et al., 2009), global fit statistics such as the minimum fit function chi-square (T_{ML}), are part of the typical output when fitting LGMs using structural equation modeling (SEM) software. With a mean structure present, Jöreskog (1967) showed that the log-likelihood function is maximized when the discrepancy function F_{ML} is minimized such that,

$$F_{ML} = \left[\ln|\hat{\Sigma}| - \ln|\Sigma| + \text{tr}\left((\Sigma - \hat{\Sigma})\hat{\Sigma}^{-1}\right) + (\mu - \hat{\mu})^T \hat{\Sigma}^{-1} (\mu - \hat{\mu}) \right], \quad (6)$$

where Σ is the observed covariance matrix of observed variables (sometimes referred to as S), $\hat{\Sigma}$ is the model-implied covariance matrix, μ is the mean vector of the observed variables (sometimes referred to as $\bar{y}$), and $\hat{\mu}$ is the model-implied mean vector (Preacher, Wichman, MacCallum, & Briggs, 2008). T_{ML} , the most common inferential statistical test for global model fit in SEM broadly, is simply calculated as $T_{ML} = (n - 1) \min(F_{ML})$. T_{ML} tests the null hypothesis $H_0 : \Sigma = \hat{\Sigma}$ and may become overpowered as sample size grows larger because trivial differences between Σ and $\hat{\Sigma}$ will result in a large test statistic value (Hu & Bentler, 1999). To address this shortcoming of T_{ML} , alternative descriptive approximate fit indices that have become widespread in the SEM literature such as the standardized root mean square residual (SRMR, although SRMR is not appropriate for growth model because it does not account for information contained in the mean structure; Wu et al., 2009), root mean square error of approximation (RMSEA), comparative fit index (CFI), or the Tucker–Lewis index (TLI; also referred as the nonnormed fit index, NNFI) have been used.

Previous Research on Small Sample Model Fit

Previous studies have addressed the properties of data–model fit criteria in the context of structural equation models generally (see, e.g., Bentler & Yuan, 1999; Ding, Velicer, & Harlow, 1995; Herzog & Boomsma, 2009; Kenny & McCoach, 2003; Marsh, Hau, Balla, & Grayson, 1998; Nevitt & Hancock, 2004; Savalei, 2010; Yuan & Bentler, 1999); however, the simulations in previous studies have yet to consider sample sizes below 100 for LGMs specifically. When sample sizes are small, T_{ML}

does not follow the appropriate χ^2 distribution and test statistics are artificially inflated, meaning that truly well-fitting models may erroneously be deemed poorly fitting (see, e.g., Bentler & Yuan, 1999; Herzog & Boomsma, 2009; Kenny & McCoach, 2003; Nevitt & Hancock, 2004). Many popular fit indices (e.g., RMSEA, CFI, TLI) include T_{ML} in the calculation, meaning that these indices will be similarly affected to varying degrees by an inflated T_{ML} statistic, making model fit particularly difficult to discern with small samples (Nevitt & Hancock, 2004).

Previous simulation studies (Fouladi, 2000; Herzog & Boomsma, 2009; Nevitt & Hancock, 2004; Savalei, 2010) have addressed this by exploring the performance of post hoc small sample corrections to T_{ML} such as those by Bartlett (1950), Swain (1975), and Yuan (2005), which have been shown to yield more appropriate rejection rates than T_{ML} in the presence of small samples. Bartlett (1950) noted that T_{ML} is suitably approximated by a χ^2 distribution with large samples but that the approximation becomes less faithful at smaller sample sizes. An exact mathematical transformation for this incongruence does not exist (see, Fujikoshi, 2000; Yuan, Tian, & Yanagihara, 2015, for a detailed discussion). Therefore, algebraic corrections for small samples are all heuristic in nature (M. W. Browne, 1982; Herzog, Boomsma, & Reinecke, 2007), each with the goal of reducing the mean of T_{ML} so that it is in line with a χ^2 distribution with the relevant degrees of freedom. These corrections operate by first estimating the model to obtain T_{ML} and then algebraically reduce T_{ML} through multiplicative post hoc corrections so that it more closely (but not exactly) follows the appropriate χ^2 distribution. The Bartlett correction (T_B) is based on the number of latent factors using an f -factor correction such that

$$T_B = \{1 - [(2\nu + 4f + 5)/6(n - 1)]\} T_{ML}, \quad (7)$$

where f is the number of latent factors and ν denotes the number of observed variables. T_B was originally intended for use with exploratory factor analysis and has been shown to perform well in such contexts (Geweke & Singleton, 1980) but tends to overcorrect when applied to SEM models (e.g., Herzog et al., 2007; Nevitt & Hancock, 2004). Yuan (2005) provided a modification to T_B in attempt to generalize the correction to a broader set of latent variables models: the Yuan correction (T_Y) is also an f -factor correction and therefore conceptually similar to T_B where

$$T_Y = \{1 - [(2\nu + 2f + 7)/6(n - 1)]\} T_{ML}. \quad (8)$$

Again, this correction was not mathematically derived and is heuristically based (Herzog & Boomsma, 2009) which Yuan himself noted in a later article, saying “this proposal [Yuan, 2005] is not statistically justified either” (Yuan et al., 2015, p. 380). Swain (1975) advanced four additional heuristic corrections, the best performing of which is referred to simply as the Swain correction (T_S) and is based entirely on degrees of freedom for the model, the number of freely estimated parameters, and sample size and is calculated by

$$T_S = \left[1 - \frac{\nu(2\nu^2 + 3\nu - 1) - q(2q^2 + 3q - 1)}{(12n)df} \right] T_{ML}, \quad (9)$$

where $q = \frac{\sqrt{1 + 4\nu(\nu + 1) - 8df} - 1}{2}$. As sample size increases, each of the three corrections approach 1 so that the asymptotic properties of T_{ML} are retained.

Fouladi (2000), Nevitt and Hancock (2004), and Herzog and Boomsma (2009) showed that T_B greatly reduced the overrejection of null hypotheses by T_{ML} with small samples although the power of T_B to identify misspecified models was reduced (i.e., it tended to overcorrect which adversely affected power) for CFA models without a mean structure and with complete data. Herzog and Boomsma (2009) found that T_S similarly reduced the tendency of T_{ML} to overreject well-fitting models but also maintained better power to reject truly misfitting models in CFA models with complete data. Fouladi (2000) advocated for T_B based on the results in her simulations which were targeted more toward nonnormality. Savalei (2010) also found that T_B and T_S versions of the Satorra–Bentler scaled T_{ML} statistic (T_{SB} ; Satorra & Bentler, 1994, 2001) perform better than the standard T_{SB} with small samples and nonnormal data. Herzog and Boomsma (2009) and Nevitt and Hancock (2004) also showed that T_Y , T_B , and T_S can be substituted for T_{ML} in equations for RMSEA, CFI, and TLI to improve the small sample performance of these indices.¹

Alternative Small Sample Methods

In addition to post hoc multiplicative corrections to T_{ML} , additional small sample test statistics have been developed (Bentler & Yuan, 1999; Yuan & Bentler, 1997, 1999). Bentler and Yuan (1999) conducted a simulation study that compared the small sample performance of T_{ML} with a variety of small sample test statistics including the residual-based asymptotic distribution-free (ADF) statistic (T_R), the Yuan and Bentler corrected version of T_R (deemed T_{YB}), the finite sample version of T_{YB} (deemed T_F), and the Satorra–Bentler test statistic (T_{SB}). Studies have recently explored the performance of methods related to these statistics with nonnormal and missing data (Yuan & Bentler, 2000; Yuan & Zhang, 2012). One drawback with the multitude of small sample statistics developed by Bentler and Yuan that were investigated in their 1999 simulation is that they are based on the ADF statistic which, although being more robust to nonnormality than T_{ML} , requires the sample size to be at least as large as the nonduplicated entries of the observed variable covariance matrix as calculated by $\nu(\nu + 1)/2$ where ν is the number of observed variables (Bentler & Yuan, 1999; Savalei, 2010). For instance, simulations and applied examples in Yuan and Bentler (2000) and Yuan and Zhang (2012) ranged from a few hundred to a few thousand which are larger than the sample sizes of interest in the current study.

For structural equation models generally, this minimum sample size for these statistics is not always inherently problematic as models typically include many latent factors, sample sizes tend to be larger than in LGMS, and the robustness to nonnormality may outweigh this drawback. However, for LGMS where nearly all variables in

the model are observed variables and samples tend to be smaller in general due to the difficulty of following individuals over extended periods of time, the minimum sample size demand to use these methods can be rather high even for relatively straightforward models with few predictors and a moderate number of repeated measures. For instance, for a model with 3 time-invariant predictors and 6 repeated measures, the sample size cannot fall below $(9 \times 10)/2 = 45$ or below 55 $(10 \times 11/2)$ for a model with 5 repeated measures and a single time-varying covariate (these values represent the number of nonredundant entries in the observed or model-implied covariance matrices). For this reason and the stated interest in small sample sizes, this article will focus on multiplicative post hoc corrections that scale T_{ML} and can be implemented regardless of sample size rather than ADF-based small sample statistics developed by Bentler and Yuan. It is, however, important to note that these methods are available and T_F in particular has been shown in previous studies to perform well with small samples, even in the face of nonnormality.

Prevalence of Small Samples With Growth Models

We have previously noted that simulation studies on LGMs have yet to include sample sizes below 100 and that ADF-based methods are unlikely to be implemented for even minimally complex models until sample sizes are in at least the high double-digits. In longitudinal studies in behavioral science research, sample sizes below 100 are quite common for reasons such as financial constraints associated with tracking participants over time, limitations of secondary data sources, or difficulty in recruiting enough participants who qualify for or are willing to participate in a study.

As supporting evidence, two large meta-analyses on aspects of personality trait changes over time by Roberts and DelVecchio (2000) and Roberts, Walton, and Viechtbauer (2006) reported that 36% (55 out of 152) and 33% (37 out of 113) of studies had sample sizes below 100, respectively. In addition, a meta-analysis of longitudinal research on prevention programs in preschools by Nelson, Westhues, and MacLeod (2003) reported that 41% (14 out of 34) of reviewed articles had fewer than 100 individuals and a meta-analysis of brain volume studies in schizophrenia patients using functional magnetic resonance imaging by Steen, Mull, McClure, Hamer, and Lieberman (2006) found that 93% (13/14) of longitudinal studies had fewer than 100 individuals. Granted, these meta-analyses come from a limited selection of subfields within the behavioral science spectrum, but assuming the prevalence of longitudinal studies with fewer than 100 individuals is roughly consistent across research areas in behavioral sciences, methodological challenges for a large proportion of studies have not been systematically addressed in the methodological literature.

Inadequacy of Small Sample Corrections With Missing Data

To outline the nature of the problem with LGMs, small samples, and missing data, we will first conduct a simulation study to demonstrate the undercorrection that

occurs in the presence of missing data based on the use of total sample size in the correction equations. Afterward, we will propose a post hoc correction for missing data inspired by the work of Rubin and Schenker (1986) in the multiple imputation literature to more appropriately incorporate the effect of missing data into the post hoc small sample corrections to T_{ML} .

Simulation Design

The number of individuals (20, 30, 50, 100), the number of repeated measures (4, 8), missing data pattern (monotone, arbitrary), and percentage of missing entries in the data matrix (0%, 10%, 20%) were manipulated in the study to explore the behavior of model fit criteria with small samples and missing data under various conditions. Percent of missing values refers to the number of cells that are missing in the data matrix. For instance, given 25 individuals and 4 repeated measures, using our definition, 10% missing would mean that 10 of the 100 cells of the data matrix had missing values. The 10% missingness condition results in roughly 80% of cases being complete, and the 20% missingness condition results in roughly 60% of cases being complete. The conditions for the percentage of missing data were chosen to roughly correspond to findings in a review by Peugh and Enders (2004) which found the mean proportion of missing data in longitudinal education studies to be about 10% with a standard deviation of about 13.

Although these corrections have been shown to perform well in previous studies with no missing data, these previous studies were restricted to covariance structure models that did not feature a mean structure that is present in latent growth models. A 0% missing condition is included to verify that the corrections are still viable in the presence of mean structure models for the limited set of conditions used in this study.

Two model conditions were generated. Model 1 featured a straightforward, linear growth model and included two binary time-invariant predictors that were generated from a standard normal distribution and dichotomized. The first predictor was dichotomized at a value of 0 yielding a 50:50 prevalence reminiscent of biological sex. The second predictor was dichotomized at 0.25 yielding a 60:40 prevalence more commonly seen in an ethnic minority status indicator. Paths from the predictors to the latent growth factors were generated such that they had a standardized effect of 0.20 on both the intercept and slope factor with both predictors together explaining approximately 10% of the total variance in intercept and slope factors, respectively. The covariance between the intercept and slope disturbances was set to be null in the population but the path was estimated in the model as would typically be done in practice since the null values would not be known a priori. Following Bauer and Curran (2003), the residual variances were chosen so that the proportion of explained variance at each time point was equal to 50%. In particular, the Θ matrix of error variances in the eight repeated measure condition had diagonal values of [1.00, 1.25, 1.75, 2.25, 3.00, 4.00, 5.25, 6.50] with all off-diagonal values being

Figure 1. Path diagram for generation of Model 1 for eight repeated measure condition. Model 2 is similar except the loadings from S to Y_j are estimated for $j \geq 3$. Displayed numbers are population values.

equal to 0 (i.e., a heterogeneous diagonal structure). The four repeated measure condition consisted only of the odd time points from the eight repeated measure condition and the model was parameterized to reflect that measurements were taken half as often. Figure 1 shows the full path diagram for Model 1 with the population values inspired by the LGM condition in Muthén and Muthén (2002).

Model 2 exhibits one of LGMs' advantages over MEMs with nonlinear growth by freely estimating the loadings from the slope factor to the repeated measure variables in what has been referred to as a latent basis model (Grimm, Ram, & Hamagami, 2011; Meredith & Tisak, 1990). In latent basis models, two slope loadings must be constrained to identify the model and set the growth scale; so the first repeated measure slope loading was constrained to 0 and the second was constrained to 1, meaning that the slope factor mean would be interpreted as the mean growth from the first to the second time point and growth at subsequent time points would be interpreted as growth from the first time point, relative to the growth between these two time points (e.g., if λ_{32} were estimated to be 2, then growth from the first to third time point would be interpreted as twice the growth from Time 1 to Time 2 in the eight repeated measure condition).²

The population values for the slope loadings were selected to be reminiscent of growth commonly seen in the learning of novel tasks or developmental science where growth is the most rapid at earlier time points but levels off as time progresses. The population values for the slope loadings were [0.00, 1.00, 3.50, 5.00, 6.00, 6.50, 6.75, 7.00] for the eight repeated measure condition. Model 2 included the same two binary time-invariant predictors as Model 1. Also similar to Model 1, the error variances had a heterogeneous diagonal structure such that the 50% of the variance in the observed repeated measures was explained at each time point, making the diagonal values of the Θ matrix equal to [1.00, 1.25, 2.75, 4.00, 5.50, 6.00, 6.25, 6.50] in the population. The path diagram for Model 2 is quite similar to Figure 1 with the exception that the slope loadings are estimated rather than constrained for observations beyond the second repeated measure.

Missing values were generated to be noninformative such that the probability of missingness was not dependent on variables excluded from the model or on the hypothetical true value itself but missingness was related to other variables included in the model (i.e., missingness would be classified as missing at random (MAR) under the classification system in Rubin, 1976). Each of the time-invariant predictor had an odds ratio of 1.60, which is typically considered to be on the border of a small and medium effect. If the odds ratio is linearly approximated via the process discussed in Chinn (2000), this would be equivalent to an r effect size of about 0.125 or a Cohen's d of 0.25. The missing data patterns were generated such that all cases had complete data for the first time point. In the monotone missingness condition, once a simulated participant missed one measurement occasion, they were missing at all subsequent measurement occasions. In the arbitrary missingness condition, generated participants could have observed values at time points after having a missing value on the previous measurement occasion.

Specifically, following Muthén and Muthén (2002), a missing data indicator was created for each time point based on a logistic regression that featured both binary variables as predictors. For the arbitrary missingness condition, the probability that a value was missing was equal for each time point, beginning at Time 2. For example, the probability that a value was missing in the 10% missingness condition, four repeated measure condition would be $\frac{0.10}{3} = 0.0\bar{3}$ at each time point (the denominator is 3 instead of 4 because the first repeated measure was always 100% complete). The monotone missingness condition was slightly more complex because missingness at each time point extended for all remaining time points. For instance, a missing value at Time 2 in the four repeated measure condition would result in three total missing values (Time 2, Time 3, and Time 4) because monotone missingness does not permit observed values to follow missing values. Because our missing data conditions were based on the percentage of elements from overall data matrix, when calculating the missing data indicator, the probability that a value was missing at each time point was weighted to reflect this such that $\text{Pr}(\text{Missingness}) = (1 - C) / \sum_{j=1}^{J-1} j$, where J is the total number of time points and C is the percent of complete entries in the data matrix. Missing values were created for each sequential time point meaning that we

Table 1. Probability That Values Were Missing at Each Time Point in the Missing Data Generation Process and the Associated Logit Value.

Percent missing	Repeated measures	Monotone		Arbitrary	
		<i>P</i> (Missing)	Logit	<i>P</i> (Missing)	Logit
10%	4	.0167	−4.072	.0333	−3.370
	8	.0036	−5.625	.0143	−4.235
20%	4	.0333	−3.370	.0667	−2.640
	8	.0072	−4.925	.0286	−3.525

first created the missing indicator for Time 2. If the indicator was 1, then we set the value of the outcome variable for Time 2 and all subsequent time points to be missing. This demonstrates why the missing data indicator at earlier time points receives more relative weight—namely, because the earlier the missing value occurs, the larger the domino-type effect on later time points becomes (e.g., with four repeated measures and monotone missingness, a generated missing value at Time 2 results in three total missing values—Time 2, Time 3, and Time 4). This must be accounted for in order to keep the overall percentage of missing data equal to the desired value. Then we created the missing indicator for observations that had observed values at Time 3 and so on. Table 1 shows the probability that a value was missing at each time point and the logit value associated with that probability.

SAS Proc Calis (Version 9.3) was used to estimate both models, output model fit criteria, and compute corrected fit criteria. FIML was used to estimate the models and accommodate missing values, conditions were fully crossed, and 2,500 replications were conducted in each cell of the design.

Results

Although more complete simulation results will be presented later on, first consider the eight repeated measure, arbitrary missingness, condition across all sample sizes as an exemplar of the results. Figure 2 shows the operating Type I error rates from the simulation with 0% missing data for T_{ML} , T_B , T_S , and T_Y . As has been shown in previous studies, T_{ML} yields highly inflated operating Type I error rates at smaller sample sizes (near 35% with 20 individuals). As has also been demonstrated in previous studies, with complete data, T_B , T_S , and T_Y all are able to correct T_{ML} such that Type I error rates are essentially at the nominal 5% rate, showing that the corrections retain their desirable performance in the presence of models that include a mean structure.

Now consider Figure 3, which presents the same conditions except that 10% (top panel) and 20% (bottom panel) of the data are missing. Even when only 10% of the data matrix is missing, T_{ML} has operating Type I error rates near 60% with 20 individuals and rejection rates still exceed 8% with 100 individuals. More important, T_B , T_S , and T_Y are less effective at correcting T_{ML} with corrected Type I error rates being

Figure 2. Type I error rates for eight repeated measures, 0% missingness, linear growth. The solid black lines represent the 5% nominal rate, dashed black lines represent 2.5% and 7.5% based on criteria in Bradley (1978) for being within reason of 5%.

near the nominal 5% rate only with 50 or more individuals. When the percentage of missing data is increased to 20%, the small sample corrections perform even worse, failing to achieve Type I error rates near the nominal level until about 100 individuals are included while the uncorrected T_{ML} statistic has rejection rates near 80% with 20 individuals and greater than 11% with 100 individuals.

From these results, it is rather clear that the small sample corrections are inadequate with missing data and the problem becomes increasingly severe as the percentage of missing data increases. To address this problem, we propose a method that scales the sample size in the small sample correction equations (Bartlett, 1950; Swain, 1975; Yuan, 2005) so that sample size more accurately reflects the amount of information as opposed to the number of people. These simulated data sets are reanalyzed using the equations that scale sample size for missing data and the results will be compared with the original equations that use total sample size.

Missing Data-Scaled Sample Size for Small Sample Corrections

With FIML, each observation contributes only what information is directly available to the log-likelihood function; however, incomplete observations do not contribute as much as complete observations. Consequently, as seen in the simulation in the previous section, using the total sample size in the correction equations undercorrects

Figure 3. Type I error rates for eight repeated measures, linear growth, 10 % arbitrary missingness (top panel), and 20% arbitrary missingness (bottom panel). The solid black lines represent the 5% nominal rate, the dashed black line represents 7.5% based on criteria in Bradley (1978) for being within reason of 5%.

because cases with missing values are weighed equally compared with cases with complete information. For example, if an individual is missing measures on seven out of eight time points, with FIML, he or she is only providing a fraction of the

information as a complete observation but this individual is counted identically to a complete observation in Equations (7) through (9).

Conceptually, this is related to (but distinct from) the problem addressed by Rubin and Schenker (1986) and expanded on by Barnard and Rubin (1999) concerning degrees of freedom for univariate inferential tests of regression coefficients with multiple imputation. Rubin and Schenker (1986) noted that degrees of freedom based on the total sample size with multiple imputation was not appropriate because it assigned equal value to directly observed and imputed values, which was attributing more information to the data than was actually obtained because the imputed values were estimated, not observed. The main idea of the Rubin–Schenker correction is to multiply the degrees of freedom for a univariate t -test by a function of data quality (one over data quality squared, more specifically), which in their correction was quantified by the fraction of missing information, or FMI, $df_{RS} = (m - 1) \left(\frac{1}{\text{FMI}^2} \right)$, where m is the number of imputations.

To account for the information lost to missing values using FIML, we propose a missing data-scaling factor that based on the logic of the Rubin–Schenker correction but with alterations so that it is applicable to the context of global data–model fit (rather than a univariate t test) and to FIML rather than multiple imputation. As discussed for the remainder of this section, the alternations required to go from the univariate context to the global model context include changing the focus from degrees of freedom to total sample size and using an alternative metric for data quality because FMI is a univariate measure.

Notationally, in Equations (7) through (9), we propose that sample size be reduced to account for missing data such that $n = M(n_{\text{Total}})$ where M is a multiplicative correction. To preserve the asymptotic properties of both T_{ML} and its small sample corrections, M must be formulated such that

- (1) as missingness $\rightarrow 0$, $M \rightarrow 1$
- (2) and $0 < M < 1$

such that M will have no impact when no missing data are present or when missing data are present but sample size is quite large.

Although Rubin and Schenker used FMI as a metric for data quality in the context of univariate degrees of freedom, there is one concern when using FMI in the context of global model fit assessment. Namely, FMI is a univariate measure—it is calculated separately for each parameter in the model. When considering global data–model fit, differing values of FMI for each parameter are difficult to reconcile. For instance, in LGMs, growth factor variances are typically more susceptible to loss of information and their FMI may be 0.60 while the FMI for factor means may be 0.10; the appropriate way to negotiate the difference between these two FMI values to accurately capture the effect missing values have on global data–model fit criteria is debatable and an appropriate method to summarize FMI globally across multiple parameters has not been addressed in the literature.

Instead, we will use the related *proportion of observed elements in the data matrix* (C) as a metric of data quality to quantify the effect of missing values. Similar to Rubin and Schenker (1986), we will use the square of data quality metric, C . Unlike Rubin and Schenker (1986), we will not take the inverse (i.e., $1/C^2$) because this will violate both Condition 1 (because the correction will approach infinity rather than 1 as missingness approaches 0) and Condition 2 (because values such as 0.50 result in a value outside the specified bounds) advanced previously. These deviations from Rubin and Schenker (1986) are attributable to their focus of degrees of freedom and our focus on a multiplicative sample size correction—as sample size grows arbitrarily large, degrees of freedom should approach ∞ whereas multiplicative corrections should approach 1 as sample size grows arbitrarily large.

FMI and C are both measures of data quality and although they are not interchangeable, Enders (2010) discussed how $1 - C$ and FMI are related conceptually with FMI typically being slightly smaller. Specifically, Enders stated “the [FMI] and the proportion of missing data [$1 - C$] are roughly equal when the variables are uncorrelated” (p. 204). Wagner (2010) demonstrated this relation between FMI and $1 - C$ in a simulation study and found that $1 - C$ and FMI for mean structure parameters remain largely equivalent until the correlation between the variables with missing values and other variables in the model exceeds about 0.30 (see figure 1 in Wagner, 2010). Importantly, as noted by Wagner (2010), C is constant across the model (rather than being unique for each parameter) which obviates the need to combine multiple values for FMI in order to best summarize the degree of information lost to missing values.

To explicitly show the location of the missing data scaling factor, the missing data-scaled, Bartlett corrected T statistic (T_{BM}) would be calculated as

$$T_{BM} = \{1 - [(2v + 4f + 5)/6(C^2(n_{Total}) - 1)]\} T_{ML}, \quad (10)$$

the missing data-scaled, Yuan corrected T statistic (T_{YM}) would be calculated as

$$T_{YM} = \{1 - [(2v + 2f + 7)/6(C^2(n_{Total}) - 1)]\} T_{ML}, \quad (11)$$

and the missing data-scaled, Swain corrected T statistic (T_{SM}) would be calculated as

$$T_{SM} = \left[1 - \frac{v(2v^2 + 3v - 1) - q(2q^2 + 3q - 1)}{(12C^2(n_{Total}))df} \right] T_{ML}. \quad (12)$$

Similar to the correction outlined in Yuan (2005) that modified the Bartlett correction, C^2 is a logical alteration that accounts for the effect of missing data on small sample corrections. The asymptotic properties of T_{ML} remain intact because as $n \rightarrow \infty$, C^2 will have increasingly less impact and the small sample correction in each of T_{BM} , T_{YM} , and T_{SM} will still approach 1. Limitations and shortcomings of such an approach are located in the “Discussion” section.

As an important note, our proposed correction is not intended to address the effect of missing data on the estimation of parameters—our correction assumes FIML has been used to accommodate missing values appropriately and the resulting parameter

estimates are unaffected by our proposed correction. Rather, our proposed correction addresses how the T statistic is adjusted to account for small samples and missing data, an issue not addressed by FIML. With FIML, although estimates are consistent provided that assumptions are met and T_{ML} calculated with FIML does incorporate information about missing values, the T statistic remains inflated when sample size is small and remains unlikely to follow the appropriate χ^2 distribution which necessitates the use of corrective procedures. Our missing data-scaling factor addresses the effect of missing data on the utility of small sample corrections, not on the estimation process or calculation of T_{ML} with FIML.

Additionally, we would like to note that the analytic scenario of interest and our proposed corrective procedure is not restricted only to LGMs—the mechanics of missing data and the issues associated with the calculation of T statistics with missing data are prevalent throughout structural equation models broadly. However, we have chosen to specifically focus on LGMs because they represent the prototypical scenario in which this precise problem will arise because of routine small sample sizes and attrition that result from the difficulty in following the same individuals over time. Our proposed method could be applied to structural equation models more generally (with or without mean structures), although the somewhat larger sample sizes typically seen in such studies may be better suited for the suite of small sample statistics proposed by Yuan, Bentler, and colleagues. However, as noted previously, small sample problems in LGMs often preclude use of these methods and render multiplicative post hoc corrections as the only available option.

Reanalyzing Simulation Data

Figure 4 partially replicates Figure 3 by showing the operating Type I error rates for correction equations that use the total sample size and also shows the results for correction equations that use missing data-scaled sample size. As seen in Figure 4, the operating Type I error rate is much closer to the nominal 5% rate when sample size is scaled to accommodate the missing data—even with as few as 20 individuals with the Bartlett correction. Because the scaling factor is proportional to the amount of missing data and to sample size, the difference between using n_{Total} and $C^2 \times n_{Total}$ shrinks both as the amount of missing data decreases and as sample size increases, preserving the asymptotic properties of both T_{ML} and the small sample corrections.

To more completely report the results of the simulation study, Table 2 shows the rejection rates for all the conditions included in the study. Because the Bartlett correction unanimously performed best in the presence of missing data, Table 2 only compares T_{ML} , T_B , and T_{BM} for clarity of exposition. Also, because arbitrary and monotone missing patterns were also quite close (within 2% across conditions), Table 2 only shows the arbitrary condition.

As Nevitt and Hancock (2004) and Herzog and Boomsma (2009) note, the corrected T statistics can be substituted in equations for approximate goodness-of-fit indices.³ Scaling sample size for missing data can also be useful for these fit indices

Figure 4. Type I error rates for eight repeated measures, linear growth, 10% arbitrary missingness (top panel), and 20% arbitrary missingness (bottom panel). The solid black lines represent the 5% nominal rate, the dashed black line represents 7.5% based on criteria in Bradley (1978) for being within reason of 5%.

as well; Figure 5 shows the median RMSEA values for the same conditions as Figures 2 through 4 based on T_{ML} and T statistics based on each of the three small sample corrections using total sample size and missing data-scaled sample size. Using the Hu and Bentler (1999) recommended cutoff⁴ for RMSEA of 0.06 (where

Table 2. Rejection Rates for T_{ML} , T_B , and T_{BM} Across All Simulation Conditions.

		Sample size											
		20			30			50			100		
		T_{ML}	T_B	T_{BM}	T_{ML}	T_B	T_{BM}	T_{ML}	T_B	T_{BM}	T_{ML}	T_B	T_{BM}
10% Missing													
4 RM													
	Model 1	15	6	5	10	6	5	9	5	5	6	5	4
	Model 2	14	5	5	10	6	6	8	5	5	6	5	5
8 RM													
	Model 1	61	13	6	32	8	4	16	6	4	9	5	4
	Model 2	63	18	10	33	10	7	17	8	6	9	5	4
20% Missing													
4 RM													
	Model 1	22	9	4	13	8	5	9	5	4	7	6	5
	Model 2	23	12	7	16	9	7	9	6	5	7	5	5
8 RM													
	Model 1	78	32	7	46	16	4	21	8	3	11	6	4
	Model 2	78	37	11	49	20	7	23	10	5	11	6	4

Note. T_{ML} = minimum fit function test statistic, T_B = Bartlett corrected test statistic using total sample size, T_{BM} = Bartlett corrected test statistics using missing data-scaled sample size, RM = repeated measures. Values in boldface indicate that T statistics had rejection rates that deviated outside the 0.025 to 0.075 range suggested by Bradley (1978) as being within reason of a 0.05 nominal rate.

lower values indicate better fit), Figure 5 shows that RMSEA values based on T_{ML} or corrections based on total sample size consider models with fewer than 50 individuals reported poor fit even though the analysis model was specified perfectly. Conversely, using the missing data-scaled sample resulted in a well-fitting model as expected, especially when using the Bartlett correction. As has been noted previously (e.g., Miles & Shevlin, 2007), CFI and TLI tend to be less affected by problems associated with T_{ML} because the index is calculated by a ratio that includes T_{ML} in both the numerator and the denominator. Although not reported, CFI and TLI did exhibit some minor problematic behavior if uncorrected with less than 50 individuals which our proposed correction was similarly able to correct. Extended results and tables regarding RMSEA, CFI, and TLI can be obtained from the authors. In results for CFI and TLI, we followed findings from Herzog and Boomsma (2009) where only the target model is adjusted while the baseline model is left unadjusted.

Applied Example

To demonstrate the utility of our proposed method, we will apply it to speech error data from Burchinal and Appelbaum (1991). These data consist of 43 children ranging in age from about 3 to 8 years. The number of speech errors made by each child was measured

Figure 5. Root mean square error of approximation (RMSEA) values for eight repeated measures, linear growth, and 20% arbitrary missingness. The solid black line represents the Hu and Bentler (1999) recommended cutoff: Values below the line indicate acceptable fit and values above the line indicate poor fit.

up to six times, approximately once per year (additional details on the data can be found in Cudeck & Haring, 2007). A plot of speech error learning curves for all 43 children is shown in Figure 6 with a superimposed mean curve included in black. As seen in Figure 6, these data show a fairly strong nonlinear association between time and speech errors made. As is common in longitudinal studies, there is also a fair amount of missingness in these data, which is mainly concentrated at later collection periods (i.e., attrition). Specifically, 68.2% of the elements of the data matrix are observed (i.e., $C = 0.682$). We will model these data in *Mplus* 7.1 using a latent basis model (because of the clear nonlinearity) and report on the T_{ML} -based data-model fit that is output by default in *Mplus* (which ignores the small sample), data-model fit based on T_B (which assumes complete data), and criteria based on the proposed T_{BM} .

Model Details

The age at which data were collected was very granular and was taken to the month. In LGMs, each possible time point must be included in the model as a separate observed variable (e.g., Biesanz et al., 2004; Hox, 2010; McNeish, 2016), meaning that there would be more possible observed variables ($6 \times 12 = 72$) than children in the data ($n = 43$) which would undoubtedly lead to convergence problems (this principle does not operate if data are modeled as an MEM; however, MEMs do not output

Figure 6. Plot of speech error learning curves for all 43 children in the Burchinal and Appelbaum (1991) data with a mean curve superimposed in black.

global data–model fit criteria). Therefore, we rounded age at the time of the data collection to the nearest whole year so that there were only six possible observed variable collection points, Age 3 through Age 8. The percent of missing data at each respective age after rounding was as follows: 2%, 9%, 9%, 35%, 63%, and 72%.

The data were then modeled as with a latent basis model such that all paths from the intercept latent variable to the observed variables were constrained to 1 while the loadings from the latent slope variable to the observed variables for Age 5 through Age 8 were freely estimated. The paths from the slope latent variable to the first and second time points (Age 3 and Age 4) were constrained to 0 and 1, respectively to give the latent variable and its associated mean an interpretable scale. The variance of the latent intercept and latent slope were estimated as was the covariance between them. The latent intercept and slope were also each predicted by an Intelligibility variable that is also included in the data ($M = 4.27$, $SD = 1.40$). The residual variances were freely estimated at each time point and a residual covariance was included between the first and second time point. The residual variance at the sixth time point was estimated to be negative, but not significantly different from 0 ($Z = -1.13$,

$p = .13$), so this residual variance was constrained to 0. The model was estimated with FIML to accommodate the missing values in *Mplus* 7.1.

Statistical Notation

The model can be written in statistical notation as

$$\text{Speech Errors}_i = \Lambda \eta_i + \mathbf{E}_i,$$

where

$$\Lambda = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & \lambda_{23} & \lambda_{24} & \lambda_{25} & \lambda_{26} \end{bmatrix}^T, \quad \eta_i = \begin{bmatrix} \text{Intercept}_i \\ \text{Slope}_i \end{bmatrix},$$

$$\mathbf{E}_i = [\varepsilon_{1i} \quad \varepsilon_{2i} \quad \varepsilon_{3i} \quad \varepsilon_{4i} \quad \varepsilon_{5i} \quad \varepsilon_{6i}]^T$$

and

$$\begin{bmatrix} \text{Intercept}_i \\ \text{Slope}_i \end{bmatrix} = \begin{bmatrix} \alpha_{\text{Int}} \\ \alpha_{\text{Slope}} \end{bmatrix} + \begin{bmatrix} \gamma_{01} \\ \gamma_{11} \end{bmatrix} [\text{Intelligibility}_i] + \begin{bmatrix} \zeta_{\text{Int},i} \\ \zeta_{\text{Slope},i} \end{bmatrix}$$

and

$$\zeta \sim MVN(\mathbf{0}, \Psi), \quad \mathbf{E} \sim MVN(\mathbf{0}, \Theta)$$

where

$$\Psi = \begin{bmatrix} \psi_{11} & & \\ \psi_{12} & \psi_{22} & \\ & & \end{bmatrix}, \quad \Theta = \begin{bmatrix} \theta_{11} & & & & & \\ \theta_{12} & \theta_{22} & & & & \\ 0 & 0 & \theta_{33} & & & \\ 0 & 0 & 0 & \theta_{44} & & \\ 0 & 0 & 0 & 0 & \theta_{55} & \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Model Results

Table 3 presents the model parameter estimates and their associated p values. Of more central interest to this article, Table 4 presents the data–model fit indices. From Table 4, it can be seen that the T_{ML} -based fit criteria (output by *Mplus*) provide evidence that the model does not fit to the data very well. T_{ML} is significant and, given the small sample size, one cannot apply the common “overpowered” argument to this model to avoid the implication of the rejected null hypothesis. Additionally, the 90% confidence interval (CI) for RMSEA is entirely greater than 0.05 and the p value for the test of close fit is less than .05.

As many previous studies and the previous simulation in this study have demonstrated, T_{ML} tends to overreject when sample sizes are small. So, after applying the

Table 3. Parameter Estimates for Latent Basis Model Fit to Speech Error Data.

Parameter	Symbol	Estimate	p value
<i>Fixed parameters</i>			
Int. mean	α_{Int}	29.10	<.001
Slope mean	α_{Slope}	-17.42	<.001
Slope loading			
Age 3	λ_{12}	0.00	—
Age 4	λ_{22}	1.00	—
Age 5	λ_{32}	1.26	<.001
Age 6	λ_{42}	1.48	<.001
Age 7	λ_{52}	1.59	<.001
Age 8	λ_{62}	1.67	<.001
Int. on intelligibility	γ_{01}	-2.37	.029
Slope on intelligibility	γ_{11}	1.44	.026
<i>Variance parameters</i>			
Var(Int.)	ψ_{11}	33.06	—
Var(Slope)	ψ_{22}	11.88	—
Cov(Int., Slope)	ψ_{12}	-19.76	.226
Var(Age3)	θ_{11}	88.15	—
Var(Age4)	θ_{22}	42.49	—
Var(Age5)	θ_{33}	14.94	—
Var(Age6)	θ_{44}	5.21	—
Var(Age7)	θ_{55}	2.16	—
Var(Age8)	θ_{66}	0.00	—
Cov(Age3, Age4)	θ_{12}	42.39	.013

Note. Int. = intercept, p values are not provided for variance parameters because they are constrained to be positive semidefinite meaning that the Z tests provided by Mplus may not be appropriate (Savalei & Kolenikov, 2008; Stram & Lee, 1994).

Table 4. Data–Model Fit Criteria Based on T_{ML} , T_B , and T_{BM} .

Criteria	T_{ML}	T_B	T_{BM}
T	29.95	26.74	22.86
p value	.018	.044	.118
RMSEA	0.144	0.126	0.101
90% RMSEA CI	[0.058, 0.222]	[0.018, 0.208]	[0.000, 0.193]
p close fit	.040	.084	.192

Note. $\nu = 7$, $f = 2$, $C = 0.682$, degrees of freedom (df) = 16, CI = confidence interval; RMSEA = root mean square error of approximation. T_{ML} = minimum fit function test statistic, T_B = Bartlett corrected test statistic using total sample size, T_{BM} = Bartlett corrected test statistics using missing data-scaled sample size. Based on M. W. Browne and Cudeck (1993), p close fit is calculated by $P = 1 - \Phi(x|\lambda^*, d)$ where $x = T$, d = degrees of freedom, $\lambda^* = 0.05^2(n - 1)d$ (a noncentrality parameter), and Φ is the cumulative distribution function of the noncentral chi-square distribution. 90% RMSEA CI for T_{ML} is part of the default output in Mplus. For, T_B and T_{BM} the 90% RMSEA CIs were calculated using the MBESS R package.

Bartlett correction, the data–model fit is slightly better but would at best be considered to have borderline acceptable fit. The p value associated with T_B is still less than .05 and the 90% CI for RMSEA straddles 0.05 but does not include 0 and the test of close fit is not significant at the .05 level but would be significant at the .10 level.

However, as argued in this article, the Bartlett correction only accounts for the small sample size were the data complete but the correction does not take missing data into consideration and thus tends to undercorrect. Using T_{BM} -based criteria, the p value associated with T_{BM} is not statistically significant at the .05 or .10 levels, the 90% CI for RMSEA straddles 0.05 (as might be expected given the imprecision associated with a sample of 43) but does include 0, and the test of close fit is not statistically significant at either the .05 or .10 levels.

Despite the fact that the parameter estimates are identical regardless of whether T_{ML} , T_B , or T_{BM} are used, each gives a very different interpretation with respect to whether the model fits the data well. If the issue of the small sample size is ignored, the T_{ML} -based criteria indicate fairly clearly that the model does not fit the data well. T_B -based criteria (that assumes complete data) resulted in better data–model fit but the interpretation for these data is not entirely clear because many of the T_B -based criteria fell very close to the cutoff points recommended in the literature. Using T_{BM} -based criteria, the data–model fit further improves compared with T_B -based criteria and T_{BM} clearly supports that the model fits the data well. Although this may seem like we are arbitrarily improving the fit of the model, keep the simulation results from Table 2 in mind—simulation results showed that T_{BM} yields appropriate Type I error rates whereas both T_{ML} and T_B had inflated rejection rates for the conditions that most closely matched the speech error data (the last row of Table 2 for the $n = 50$ column).

Note that in Table 4, the RMSEA values appear to indicate somewhat poor fit across conditions and RMSEA does not agree with the inferential decision from T_{BM} . Although this intuitively seems problematic, issues with RMSEA in models with few degrees of freedom and small samples has been extensively discussed in Kenny, Kaniskan, and McCoach (2015). Kenny et al. (2015) found that rejection rates for perfectly specified models climbed as both the model degrees of freedom and sample size decreased. In their simulation study, with a sample size of 50 and 16 degrees of freedom (very closely matching the applied example at hand), nearly 15% of perfectly specified models were rejected based on a cutoff of 0.05 (as opposed to 0% rejection with sample sizes of 200 or higher with 16 degrees of freedom). This further demonstrates the utility of the small sample methods we are proposing—with small samples, inferential tests are the unequivocal best option for assessing data–model fit, so it is vital to ensure that the p values and resulting inferential decisions can be trusted.

Discussion

Limited previous research on post hoc small sample corrections to T_{ML} and missing data found inflated Type I error rates with samples below 100 which was further corroborated by the simulation performed in this study in the previously unstudied

context of LGMs. A post hoc, missing data-scaling factor for the sample size in the small sample correction equations with FIML was found to provide much better Type I error rates and improved performance of the approximate goodness-of-fit indices under a variety of conditions including monotone and arbitrary missing data patterns when the missingness was MAR. Based on the simulations in this study, the Bartlett correction with the missing data-scaling factor is recommended for models with small samples and missing data treated with FIML and the Yuan correction is recommended for models with small samples and complete data. The missing data-scaled post hoc corrections maintained satisfactory performance for LGMs with 20% missing data (60% complete cases) with as few as 20 total individuals.

Practically speaking, researchers may be interested at which point a sample size is small enough to be considered a “small sample problem.” In the models used in the simulation, the point was somewhere between 50 and 100; however, note that as model complexity increases, small sample issues occur at larger and larger samples (e.g., McNeish & Stapleton, 2016). For example, in the models in the simulation, Type I error rates were more inflated for the latent basis model compared with the linear growth model and also for the eight repeated measure model compared with the four repeated measure model (the residuals followed a heterogeneous structure so more repeated measures required more estimated parameters). For models with more complicated growth trajectories or several time-invariant or time-varying predictors, small sample issues will be present at higher sample sizes so an exact cutoff of what exactly constitutes a small sample cannot be definitively stated.

Fortunately, each of the three post hoc small sample corrections with missing data-scaled sample size proposed in this article preserve the asymptotic properties of T_{ML} because the correction approaches 1 as sample size increases and/or missing data decreases. This means that researchers do not have to explicitly decide when to or when not to use these methods unless it is abundantly clear that the sample size is sufficiently large to avoid the “small sample” classification. Even if the amount of missing data is fairly sizeable, a large sample size will obviate the effect of the missing data scaling factor and the post hoc small sample correction will still approach 1. A similar mechanism will apply if one has the reverse situation of a small sample and few missing values. Thus, if one has an adequate sample size, both the post hoc small sample correction and the associated missing data-scaling factor will essentially have no effect and their use will not adversely affect the results.

The small sample corrections under investigation in this study can also be applied to robust statistic as well such as the Yuan–Bentler $T2^*$ (Yuan & Bentler, 2000; commonly referred to by its *Mplus* designation as “MLR”) or the Satorra–Bentler test statistic (Satorra & Bentler, 1994, 2001). Savalei (2010) discussed how these test statistics are also asymptotically chi-square distributed, so the multiplicative small sample correction factors directly apply. To extend her logic one step forward, the missing data correction proposed in this article could similarly apply to the inferential tests produced by these robust estimators. We did not study this explicitly in the

current study and further studies would be needed to assess the effectiveness of the proposed correction with robust test statistics.

As with all studies, this one had certain limitations. Ordinarily, it would be best practice to follow a study of Type I error rates with a power study that fits misspecified models and assesses the extent to which the statistic(s) under investigation can identify the misspecified models. From the results of this study, a power study of this type was not warranted because only one method (T_{BM}) controlled the Type I error rate. A comparison of the relative power of different methods cannot be conducted because power is uninterpretable as the Type I error rate is not well controlled. Additional studies on empirical power would be needed for alternative contexts (e.g., different model types) in which multiple methods are able to control Type I error rates.

Second, similar to much of the literature in this area, the correction we proposed was based on heuristic grounds. As we noted in this article, to accommodate the effect of missing data, the small sample corrections must be altered so that information about data quality is included. We chose to use C , the percentage of complete values, because Wagner (2010) had noted the inherent advantage of this metric when considering the model globally. Although our decisions are based on prior research, we nonetheless are fully aware that corrections based on alternative choices could also perform well or even better than our proposed correction. For instance, future work could consider using FMI instead of C if a method to combine parameter-specific FMI values into a global summary could be devised and defended (perhaps averaging FMI over all parameters is sufficient). This could be advantageous if missingness is highly related to missing values as Wagner (2010) showed that FMI and C are the most disparate under such conditions. Alternatively, although there is a precedent for squaring the metric of data quality (i.e., FMI or C) from Rubin and Schenker (1986), this is not to say that this is the only choice and other functions may be better suited for such a purpose. Needless to say, research on methods for assessing fit for models that simultaneously have small samples and missing data is needed to help answer these questions.

Relatedly, as currently proposed, the C^2 correction may be deficient when there are very high correlations (i.e., > 0.70) between missing values and other variables in the model. As noted in Wagner (2010) and Graham (2012), the value of C and 1-FMI are identical in the case of MCAR data but they diverge as the MAR assumption becomes increasingly strong. When there are strong relations between hypothetical missing values and those observed on other variables, the observed variables can account for a portion of the variance in the data that would have been present if values were collected (i.e., if the multiple correlation of the missing values with observed values is 0.50, 25% of the information from the missing values can be gleaned from the observed information). Because of this, in the context of strong correlations between hypothetical missing values and observed variables, the proposed correction may tend to overcorrect (i.e., deflate the Type I error rate) because C will become a less accurate approximation for the effect of missing data and will overestimate the effect of missing data (i.e., it will not account for the portion of the

missing values that can be accounted for by other observed information). The data generation process in our simulation study where odds ratios between observed variables and missing values were on the border of a small and medium did not feature relations that were strong enough to observe this theoretical behavior.

Third, this study focused on LGMs which form only a small subset of the broader set of structural equation models. As LGMs are essentially CFA models with an imposed mean structure, it seems intuitively reasonable that scaling sample size based on missing data would work reasonably well for CFA models generally. As noted previously, we restricted our focus to LGMs because these models are the most likely to feature both reduced sample sizes and missing data due to the inherent difficulty of repeatedly measuring the same individuals over time and the unavoidable attrition that occurs in many studies. Additional studies could investigate the performance of the missing data-scaled sample size equations to different types of models.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. The `swain` and `FAiR` packages in R can perform the Swain correction and the `FAiR` package can also compute the Bartlett correction. These corrections are not available in commercial software although they can be calculated by hand or with a spreadsheet without much effort. An Excel spreadsheet for calculating corrections is provided on the first author's personal website at <https://sites.google.com/site/danielmmcneish/academic-work/smallsamplecorrections>
2. Constrained loadings are typically chosen to correspond to a substantively important time-frame so that the interpretation is substantively relevant; however, since the data in the simulation are artificial, the constrained loadings were selected rather arbitrarily.
3. Note that with the small sample sizes of interest in this article, the need to consider approximate goodness-of-fit indices is reduced because the researchers cannot invoke the argument that the T statistic will be overpowered.
4. Recommendations from the Hu and Bentler studies were not directly intended for LGMs and the use of cutoffs and the approximate fit indices in general has been a recent point of contention (see, e.g., Barrett, 2007; Hayduk, Cummings, Boadu, Pazderka-Robinson, & Boulianne, 2007). The practice of overgeneralizing Hu and Bentler's guidelines has been previously noted (Marsh, Hau, & Wen, 2004).

References

- Barnard, J., & Rubin, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86, 948-955.

- Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, 42, 815-824.
- Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Statistical Psychology*, 3, 77-85.
- Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods*, 8, 338-363.
- Bell, B. A., Morgan, G. B., Schoeneberger, J. A., Kromrey, J. D., & Ferron, J. M. (2014). How low can you go? *Methodology*, 10, 1-11.
- Bentler, P. M., & Yuan, K. H. (1999). Structural equation modeling with small samples: Test statistics. *Multivariate Behavioral Research*, 34, 181-197.
- Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, NJ: Wiley.
- Biesanz, J. C., Deeb-Sossa, N., Papadakis, A. A., Bollen, K. A., & Curran, P. J. (2004). The role of coding time in estimating and interpreting growth curve models. *Psychological Methods*, 9, 30-52.
- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, 144-152.
- Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in applied multivariate analysis* (pp. 72-142). Cambridge, England: Cambridge University Press.
- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.
- Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, 1, 473-514.
- Burchinal, M., & Appelbaum, M. I. (1991). Estimating individual developmental functions: Methods and their assumptions. *Child Development*, 62, 23-43.
- Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta analysis. *Statistics in Medicine*, 19, 3127-3131.
- Chou, C. P., Bentler, P. M., & Pentz, M. A. (1998). Comparisons of two statistical approaches to study growth curves: The multilevel model and the latent curve analysis. *Structural Equation Modeling*, 5, 247-266.
- Cudeck, R., & Haring, J. R. (2007). Analysis of nonlinear patterns of change with random coefficient models. *Annual Review of Psychology*, 58, 615-637.
- Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research*, 38, 529-569.
- Ding, L., Velicer, W. F., & Harlow, L. L. (1995). Effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modeling fit indices. *Structural Equation Modeling*, 2, 119-143.
- Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.
- Fouladi, R. T. (2000). Performance of modified test statistics in covariance and correlation structure analysis under conditions of multivariate nonnormality. *Structural Equation Modeling*, 7, 356-410.
- Fujikoshi, Y. (2000). Transformations with improved chi-squared approximations. *Journal of Multivariate Analysis*, 72, 249-263.
- Geweke, J. F., & Singleton, K. J. (1980). Interpreting the likelihood ratio statistic in factor models when sample size is small. *Journal of the American Statistical Association*, 75, 133-137.

- Graham, J. W. (2012). *Missing data: Analysis and design*. New York, NY: Springer.
- Grimm, K. J., Ram, N., & Hamagami, F. (2011). Nonlinear growth curves in developmental research. *Child Development*, 82, 1357-1371.
- Hayduk, L., Cummings, G., Boadu, K., Pazderka-Robinson, H., & Boulianne, S. (2007). Testing! testing! one, two, three—Testing the theory in structural equation models! *Personality and Individual Differences*, 42, 841-850.
- Herzog, W., & Boomsma, A. (2009). Small-sample robust estimators of noncentrality-based and incremental model fit. *Structural Equation Modeling*, 16, 1-27.
- Herzog, W., Boomsma, A., & Reinecke, S. (2007). The model-size effect on traditional and modified tests of covariance structures. *Structural Equation Modeling*, 14, 361-390.
- Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.
- 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.
- Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, 32, 443-482.
- Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, 44, 486-507.
- Kenny, D. A., & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling*, 10, 333-351.
- Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, 1, 86-92.
- Marsh, H. W., Hau, K. T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, 33, 181-220.
- Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, 11, 320-341.
- McNeish, D. (2016). Using data-dependent priors to mitigate small sample size bias in latent growth models: A discussion and illustration using *Mplus*. *Journal of Educational and Behavioral Statistics*, 41, 27-56.
- McNeish, D., & Stapleton, L. M. (2016). The effect of small sample size on two level model estimates: A review and illustration. *Educational Psychology Review*, 28, 295-314.
- Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, 55, 107-122.
- Miles, J., & Shevlin, M. (2007). A time and a place for incremental fit indices. *Personality and Individual Differences*, 42, 869-874.
- Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, 9, 599-620.
- Nelson, G., Westhues, A., & MacLeod, J. (2003). A meta-analysis of longitudinal research on preschool prevention programs for children. *Prevention & Treatment*, 6, 31a.
- Nevitt, J., & Hancock, G. R. (2004). Evaluating small sample approaches for model test statistics in structural equation modeling. *Multivariate Behavioral Research*, 39, 439-478.
- Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, 74, 525-556.
- Preacher, K. J., Wichman, A. L., MacCallum, R. C., & Briggs, N. E. (2008). *Latent growth curve modeling*. Thousand Oaks, CA: Sage.

- Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin*, 126, 3-25.
- Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin*, 132, 1-25.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- Rubin, D. B., & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.
- Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399-419). Thousand Oaks, CA: Sage.
- Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, 66, 507-514.
- Savalei, V. (2010). Small sample statistics for incomplete nonnormal data: Extensions of complete data formulae and a Monte Carlo comparison. *Structural Equation Modeling*, 17, 241-264.
- Savalei, V., & Kolenikov, S. (2008). Constrained versus unconstrained estimation in structural equation modeling. *Psychological Methods*, 13, 150-170.
- Steen, R. G., Mull, C., McClure, R., Hamer, R. M., & Lieberman, J. A. (2006). Brain volume in first-episode schizophrenia: Systematic review and meta-analysis of magnetic resonance imaging studies. *British Journal of Psychiatry*, 188, 510-518.
- Stram, D. O., & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, 50, 1171-1177.
- Swain, A. J. (1975). *Analysis of parametric structures for variance matrices* (Unpublished doctoral dissertation). Department of Statistics, University of Adelaide, Adelaide, Australia.
- Wagner, J. (2010). The fraction of missing information as a tool for monitoring the quality of survey data. *Public Opinion Quarterly*, 74, 223-243.
- Wu, W., West, S. G., & Taylor, A. B. (2009). Evaluating model fit for growth curve models: Integration of fit indices from SEM and MLM frameworks. *Psychological Methods*, 14, 183-201.
- Yuan, K.-H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research*, 40, 115-148.
- Yuan, K.-H., & Bentler, P. M. (1997). Mean and covariance structure analysis: Theoretical and practical improvements. *Journal of the American Statistical Association*, 92, 767-774.
- Yuan, K.-H., & Bentler, P. M. (1999). F tests for mean and covariance structure analysis. *Journal of Educational and Behavioral Statistics*, 24, 225-243.
- Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, 30, 167-202.
- Yuan, K.-H., Tian, Y., & Yanagihara, H. (2015). Empirical correction to the likelihood ratio statistic for structural equation modeling with many variables. *Psychometrika*, 80, 379-405.
- Yuan, K.-H., & Zhang, Z. (2012). Robust structural equation modeling with missing data and auxiliary variables. *Psychometrika*, 77, 803-826.