Psychological Methods
2018, Vol. 23, No. 1, 76–93

1082-989X/18/$12.00 http://dx.doi.org/10.1037/met0000102

Assessing the Fit of Structural Equation Models With Multiply Imputed Data

Craig K. Enders and Maxwell Mansolf
University of California, Los Angeles

Multiple imputation has enjoyed widespread use in social science applications, yet the application of imputation-based inference to structural equation modeling has received virtually no attention in the literature. Thus, this study has 2 overarching goals: evaluate the application of Meng and Rubin's (1992) pooling procedure for likelihood ratio statistic to the SEM test of model fit, and explore the possibility of using this test statistic to define imputation-based versions of common fit indices such as the TLI, CFI, and RMSEA. Computer simulation results suggested that, when applied to a correctly specified model, the pooled likelihood ratio statistic performed well as a global test of model fit and was closely calibrated to the corresponding full information maximum likelihood (FIML) test statistic. However, when applied to misspecified models with high rates of missingness (30%–40%), the imputation-based test statistic generally exhibited lower power than that of FIML. Using the pooled test statistic to construct imputation-based versions of the TLI, CFI, and RMSEA worked well and produced indices that were well-calibrated with those of full information maximum likelihood estimation. This article gives Mplus and R code to implement the pooled test statistic, and it offers a number of recommendations for future research.

Keywords: multiple imputation, missing data, model fit, structural equation modeling

The last two decades have seen a dramatic increase in the application of missing data handling methods that assume a missing at random (MAR) mechanism, a condition where the values of *Y* carry no information about the probability of missing data on *Y* after conditioning on observed variables in the analysis (or imputation) model (Little & Rubin, 2002; Rubin, 1976). To date, full information maximum likelihood (FIML) estimation and multiple imputation are the principal MAR-based approaches in social science applications; FIML estimates model parameters directly from the observed data, whereas multiple imputation fills in the data prior to analysis. Given a common set of input variables and reasonable sample size, the methods tend to produce equivalent estimates with multivariate normal data (Collins, Schafer, & Kam, 2001; Meng, 1994; Schafer, 2003), and there is often little reason to prefer one technique to the other; differences may emerge at very small sample sizes or with nonnormal data, however (von Hippel, 2015; Yuan, Yang-Wallentin, & Bentler, 2012).[1]

Statistical inference is one area where FIML and imputation can diverge. With complete data, researchers routinely use Wald or likelihood ratio chi-square tests to evaluate multiparameter hypotheses. The FIML versions of these tests statistics are natural extensions of their complete-data counterparts, whereas the imputation-based tests require specialized pooling procedures (Li, Raghunathan, & Rubin, 1991; Meng & Rubin, 1992; Reiter & Raghunathan, 2007; Rubin, 1987a). A good deal of recent research has focused on FIML estimation and inference, and the properties of FIML test statistics are generally well understood (Reiter, 2007; Yuan, 2009; Yuan & Bentler, 2000, 2010; Yuan & Savalei, 2014; Yuan & Zhang, 2012). In contrast, much less is known about the corresponding test statistics for multiply imputed data. Much of the research to date has focused on improving the small-sample properties of single-parameter *t* statistics (Barnard & Rubin, 1999; Reiter, 2007; Steele, Wang, & Raftery, 2010), and very few studies have examined Wald and likelihood ratio tests (Lee & Cai, 2012; Liu & Enders, 2016).

The application of multiple imputation inference to structural equation models (SEMs) is an important issue that has received virtually no attention in the literature. To date, Lee and Cai (2012) is the only study to consider this problem. These authors propose a two-stage approach that is analogous to two-stage FIML estimation (Savalei & Bentler, 2009; Savalei & Falk, 2014; Savalei & Rhemtulla, 2014; Yuan, Tong, & Zhang, 2014). Following imputation, researchers first estimate a saturated model and apply Rubin's (1987a) pooling rules to the variances and covariances. The resulting covariance matrix then serves as input data for a standard SEM analysis. The complete-data fitting function yields invalid test statistics and standard errors with this approach, but Lee and Cai (2012) derive a residual-based test statistic that appears to exhibit good performance. Although the authors do not examine other fit indices, they suggest that the residual-based chi-square can be used to compute indices such as the Tucker-Lewis Index (TLI; also known as the non-normed fit index, or

This article was published Online First November 28, 2016.

Craig K. Enders and Maxwell Mansolf, Department of Psychology, University of California, Los Angeles.

An earlier version of this work was presented at the 2015 meeting of the Society of Multivariate Experimental Psychology.

Correspondence concerning this article should be addressed to Craig K. Enders, Department of Psychology, University of California, Los Angeles, Box 951963, Los Angeles, CA 90095-1563. E-mail: cenders@psych.ucla.edu

[1] It is difficult to define "reasonable," but Yuan et al. (2012) found the largest differences between maximum likelihood and multiple imputation at sample sizes of 100 or smaller.

NNFI; Bentler & Bonett, 1980; Tucker & Lewis, 1973), comparative fit index (CFI; Bentler, 1990), and root mean square error of approximation (RMSEA; Browne & Cudeck, 1993; Steiger, 1989, 1990; Steiger & Lind, 1980), among others. Currently, the two-stage approach is available only as a SAS macro that works in conjunction with the CALIS procedure. For most researchers, the two-stage procedure is prohibitively difficult to implement in other software packages.

Meng and Rubin's (1992) pooling procedure for likelihood ratio chi-square statistics is a second and yet-unstudied option for assessing model fit with multiply imputed data. As explained later in the article, the pooled test statistic requires two passes through the data. The first step fits a pair of models to each imputed data set and averages the resulting parameter estimates and model fit statistics. The mean likelihood ratio test is then computed a second time after estimating the models with their parameters fixed at the pooled estimates. The final test statistic is a function of the average likelihood ratio values from the two runs. Meng and Rubin's test appears to provide adequate Type I error control and comparable power to FIML with multiple regression models (Liu & Enders, 2016), but it is yet unknown whether the procedure is suitable for evaluating SEM fit. As noted later, the application of the Meng and Rubin's (1992) procedure to SEM raises interesting issues that are not germane to standard regression models. The pooled likelihood ratio statistic is currently available in Mplus (Muthén & Muthén, 1998–2012) and the semTools package in R (Pornprasertmanit et al., 2016).

Because the properties of the FIML test statistic are generally well understood, it is reasonable to ask why one would choose multiple imputation for an SEM analysis, particularly given that FIML is so readily available in software. There are a number of compelling reasons. First, imputation is well suited for mixtures of incomplete categorical and continuous variables, whereas the implementation of FIML in popular software packages generally requires multivariate normality. As a simple example, consider a MIMIC model with an incomplete categorical covariate (e.g., a set of incomplete dummy codes). Existing SEM software packages would force the user to specify a normal distribution for such variables in order to retain the incomplete cases, but doing so could compromise the quality of the resulting estimates. In contrast, categorical imputation schemes based on logistic or probit regression are widely available in software packages (Carpenter & Kenward, 2013; Enders, Mistler, & Keller, 2016; van Buuren, 2012), and researchers can easily generate imputations that honor a diverse set of metrics. A second related issue occurs with questionnaire items where researchers often form item parcels before including such variables in an SEM (Little, Rhemtulla, Gibson, & Schoemann, 2013). FIML provides no mechanism for computing such a composite when the parcel components are incomplete, whereas this problem is trivial to deal with in the imputation framework. Here again, a researcher could first use a categorical imputation scheme to fill in the missing item responses, and then compute parcels for the SEM analysis. A third reason to prefer imputation is the ease with which it can accommodate auxiliary variables. Although FIML can also incorporate auxiliary variables with Graham's (2003) saturated correlates approach, it is well known that this method suffers from convergence problems because it imposes an illogical structure on certain covariance matrices (Savalei & Bentler, 2009). Finally, the standard independence model that posits uncorrelated variables necessarily requires a missing completely at random (MCAR) mechanism because the absence of correlation precludes the conditioning necessary to satisfy MAR. Whether this feature materially impacts FIML fit indices is an open question, but it is nevertheless a limitation.

Given the widespread application of SEMs in the behavioral science literature, it is important to develop and evaluate methods for assessing model fit with multiply imputed data. As such, this study has two primary goals: evaluate the application of Meng and Rubin's (1992) pooling procedure to the chi-square test of model fit, and explore the possibility of using the test statistic to derive imputation-based versions of common fit indices such as the TLI, CFI, and RMSEA. It is important to consider Meng and Rubin's procedure in the context of SEM because it is subtly different from the complete-data (or FIML) fit statistic. For example, it is widely known that the likelihood ratio test is invariant to changes in model parameterization or identification, such that alternate specifications of the same model (e.g., identifying a confirmatory factor analysis model by fixing either a loading or the factor variance to unity) will produce the same fit statistic (Gonzalez & Griffin, 2001). Because Meng and Rubin's test does not share this property (Schafer, 1997), arbitrary identification constraints can potentially impact model fit. Whether this feature of the test statistic has any material impact on SEM analyses is an open question.

Currently, the literature offers no guidance on constructing SEM fit indices with multiply imputed data. Rubin's (1987a) pooling rules apply only to parameters and assume that, over repeated samples, estimates are normally distributed around some population value (van Buuren, 2012; White, Royston, & Wood, 2011). Because the distributions of most SEM fit indices are either unknown or not normal (Bentler, 1990; Bollen & Stine, 1992; Browne & Cudeck, 1992), averaging these quantities in the same manner as estimates is at odds with multiple imputation theory.[2] Several fit indices that currently enjoy widespread use are simple functions of the likelihood ratio test, including the CFI, TLI, and RMSEA, among others. A simple yet unstudied strategy for constructing fit indices with multiply imputed data is to substitute Meng and Rubin's (1992) test statistic into standard complete-data fit formulas. Lee and Cai (2012) raised this possibility with their residual-based chi-square test but did not pursue the issue. We explore this strategy later in the article.

The organization of this article is as follows. First, we give a brief review of multiple imputation, focusing primarily on Rubin's (1987a) pooling rules and single-parameter inference. Second, we give a detailed overview of the Wald and likelihood ratio tests for multiply imputed data (Li et al., 1991; Meng & Rubin, 1992; Reiter & Raghunathan, 2007; Rubin, 1987a) and discuss the application of these tests to SEM. Third, we use computer simulations to study the behavior of Meng and Rubin's (1992) test statistic with correctly and incorrectly specified structural equation models. Fourth, we use Meng and Rubin's test to define imputation-based versions of the CFI, TLI, and RMSEA, and we use computer simulations to investigate these indices. Finally, we conclude with a brief data analysis example and discussion.

[2] The notable exception is the standardized root mean residual (SRMR), which is amenable to pooling because it compares two estimates of the population covariance matrix.

Multiple Imputation

Multiple imputation involves three major steps. In the first step, the researcher creates several copies of the incomplete data (e.g., $m = 20$ or more; Graham, Olchowski, & Gilreath, 2007), each of which contains different plausible replacement values. Imputation typically employs a regression-based procedure in the Bayesian framework. Because the imputation process itself is not central to this article, we refer readers to other sources for additional information (Enders, 2010; Graham, 2012; Schafer, 1997; Sinharay, Stern, & Russell, 2001; van Buuren, 2012). After creating the imputed data sets, the researcher then performs a statistical analysis (e.g., an SEM) on each filled-in data set. In the final step, the m estimates of each parameter are pooled using a simple arithmetic average

$$\bar{\theta} = \frac{1}{m}\sum_{t=1}^{m} \theta^{(t)} \tag{1}$$

where $\bar{\theta}$ is the pooled point estimate, and $\theta^{(t)}$ is the estimate from data set t.

Following the pooling phase, researchers routinely use an imputation-based version of the familiar Wald test (Buse, 1982) to evaluate individual parameter estimates. Although we focus strictly on model fit in this manuscript, we briefly describe the Wald statistic here because Meng and Rubin (1992) used this test as the basis for deriving their pooling procedure for likelihood ratio statistics. As such, understanding the composition of the Wald test provides insight into the likelihood ratio statistic. To illustrate the Wald test, consider a k-element vector of pooled estimates $\bar{\boldsymbol{\theta}}$ and a corresponding vector of hypothesized values $\boldsymbol{\theta}_0$ (e.g., a zero vector). The Wald statistic includes a complete-data part and an adjustment term that accounts for missing data uncertainty. The former component mimics the structure of the complete-data Wald test

$$(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \bar{\mathbf{V}}^{-1} (\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \tag{2}$$

where $\bar{\mathbf{V}}$ is the k by k matrix of sampling variances and covariances (i.e., the within-imputation covariance matrix) computed by averaging parameter covariance matrices from the m complete data sets, as follows.

$$\bar{\mathbf{V}} = \frac{1}{m}\sum_{t=1}^{m} \mathbf{V}^{(t)} \tag{3}$$

Equation 2 is an inappropriate test statistic because $\bar{\mathbf{V}}$ reflects only complete-data sampling error and thus underestimates uncertainty about $\bar{\boldsymbol{\theta}}$. This problem is addressed by attenuating Equation 2 to compensate for missing data noise. This adjustment requires the variance-covariance matrix of the estimates across the m data sets (i.e., the between-imputation covariance matrix)

$$\mathbf{B} = \frac{1}{m}\sum_{t=1}^{m} (\boldsymbol{\theta}^{(t)} - \bar{\boldsymbol{\theta}})(\boldsymbol{\theta}^{(t)} - \bar{\boldsymbol{\theta}})^T \tag{4}$$

and the average relative increase in variance

$$\bar{r} = (1 + m^{-1})\mathrm{tr}(\mathbf{B}\bar{\mathbf{V}}^{-1})/k \tag{5}$$

a quantity that expresses the average amount of missing data uncertainty in $\mathbf{B}$ as a proportion of the complete-data sampling variances in $\bar{\mathbf{V}}$ (Li et al., 1991; Rubin, 1987b).

Finally, the imputation-based Wald test is computed as follows

$$W = \frac{(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \bar{\mathbf{V}}^{-1} (\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)}{1 + \bar{r}} \tag{6}$$

and a probability value for the test is obtained by referencing W to a chi-square distribution with k degrees of freedom (Asparouhov & Muthén, 2010) or by referencing W/k to the approximate F reference distribution given by (Li et al., 1991). Equation 6 illustrates that the average relative increase in variance operates as a correction factor that attenuates the pooled components in the numerator to compensate for missing data uncertainty. Specifically, note that $\bar{r}$ reaches its theoretical minimum value of zero when all elements in $\mathbf{B}$ equal zero, a situation that can only occur with no missing data. In this case, the denominator of Equation 6 equals unity and the expression reduces to the usual complete-data Wald test. In contrast, $\bar{r}$ increases as the missing data rates increase, and the test statistic shrinks proportionately to compensate.

The relative increase in variance—which relies only on the diagonal elements of $\mathbf{B}\bar{\mathbf{V}}^{-1}$—was originally developed to address the fact that $\mathbf{B}$ is very noisy and may not be full rank (Meng & Rubin, 1992; Reiter & Raghunathan, 2007; Schafer, 1997; van Buuren, 2012). Because the average relative increase (a scalar quantity) replaces the matrix of ratios in $\mathbf{B}\bar{\mathbf{V}}^{-1}$, the Wald test assumes that the elements in $\mathbf{B}$ are proportional to their corresponding elements in $\bar{\mathbf{V}}$, such that missing data uniformly increase all sampling variances and covariances by the same relative amount (e.g., the missing data increase the squared standard errors of all parameters by 15%). In a set of orthogonal variables, $\mathbf{B}$ and $\bar{\mathbf{V}}$ are proportional when all variables have the same missing data rate, but the proportional increase in noise from missing data generally depends on the correlations among the variables as well. Limited simulation results suggest that moderate violations of proportionality are not problematic (Li et al., 1991), and an alternative procedure that uses the entire $\mathbf{B}$ matrix to adjust the test statistic does not produce suitable inferences, at least in multiple regression models (Liu & Enders, 2016). We highlight this issue because the likelihood ratio test described in the next section also assumes that $\mathbf{B}$ and $\bar{\mathbf{V}}$ differ by a proportional constant, although the reasons behind the assumption are much less obvious.

Pooled Likelihood Ratio Test

Like the Wald test, the likelihood ratio statistic can be used to evaluate individual parameters (or sets of parameters) or to evaluate global model fit. We focus on the latter issue in this article. The likelihood ratio statistic has the researcher fit a pair of nested models, the first of which is a more general model, the second of which is a more restrictive model that places constraints on certain parameters. In SEM applications, the general model is often saturated, and the restrictive model reflects the researcher's hypotheses. In addition to the global test of model fit, the incremental fit indices described later in the paper also require a likelihood ratio test comparing the saturated model with a baseline or independence model that posits no correlation among the variables.

Meng and Rubin's (1992) pooling procedure for likelihood ratio tests requires two passes through the data. The first pass fits the models of interest to each imputed data set and averages the m likelihood ratio statistics, as follows

$$\overline{T} = \frac{1}{m}\sum_{t=1}^{m} -2l(\boldsymbol{\theta}_0^{(t)} \mid \mathbf{Y}^{(t)}) + 2l(\boldsymbol{\theta}_1^{(t)} \mid \mathbf{Y}^{(t)}) \tag{7}$$

where $\overline{T}$ is the average likelihood ratio test, $\mathbf{Y}^{(t)}$ denotes the data from imputation t, $\boldsymbol{\theta}_1^{(t)}$ and $\boldsymbol{\theta}_0^{(t)}$ are estimates from the general and restrictive (e.g., saturated and hypothesized) models, respectively, in data set t, and $(\boldsymbol{\theta}_1^{(t)} \mid \mathbf{Y}^{(t)})$ and $l(\boldsymbol{\theta}_0^{(t)} \mid \mathbf{Y}^{(t)})$ are the corresponding log likelihood values. The second pass computes the average likelihood ratio test from models with parameters fixed at their pooled values. The average likelihood ratio statistic is

$$\overline{T}' = \frac{1}{m}\sum_{t=1}^{m} -2l(\bar{\boldsymbol{\theta}}_0 \mid \mathbf{Y}^{(t)}) + 2l(\bar{\boldsymbol{\theta}}_1 \mid \mathbf{Y}^{(t)}) \tag{8}$$

where $\bar{\boldsymbol{\theta}}_1$ and $\bar{\boldsymbol{\theta}}_0$ are the pooled estimates from the general and restrictive models, respectively, and $l(\bar{\boldsymbol{\theta}}_1 \mid \mathbf{Y}^{(t)})$ and $l(\bar{\boldsymbol{\theta}}_0 \mid \mathbf{Y}^{(t)})$ are the corresponding log likelihood values from data set t evaluated at the pooled estimates. Equation 8 can be obtained analytically by first using the pooled estimates to compute the model-implied mean vector and covariance matrix, then evaluating the log likelihood of each imputed data set at these quantities. Alternatively, the researcher could refit the two models to each data set with estimates fixed at their pooled values.

After obtaining the two averages, the pooled likelihood ratio statistic is

$$T_{(\text{imp})} = \frac{\overline{T}'}{1 + \bar{r}_T} \tag{9}$$

where

$$\bar{r}_T = \frac{m+1}{k(m-1)}(\overline{T} - \overline{T}') \tag{10}$$

is an alternate estimate of the average relative increase in variance. A probability value is obtained by referencing $T_{(\text{imp})}$ to a chi-square distribution with k degrees of freedom (Asparouhov & Muthén, 2010) or by referencing $T_{(\text{imp})}/k$ to the approximate F distribution given by (Li et al., 1991). In line with most SEM applications, we are primarily interested in whether the test statistic approximates a chi-square variate.

Notice that the pooled likelihood ratio statistic shares the same form as the Wald test from Equation 6, with a component that depends on the pooled estimates, and an adjustment term that attenuates the test statistic to account for missing data uncertainty. In fact, Meng and Rubin (1992) relied on the asymptotic equivalence of the Wald and likelihood ratio tests to derive $T_{(\text{imp})}$, starting with the two components of the Wald test and developing analogous quantities in the likelihood framework. Specifically, the authors show the following asymptotic equivalencies.

$$\begin{aligned} \overline{T}' &\cong (\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \overline{\mathbf{V}}^{-1} (\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ \frac{m+1}{k(m-1)}(\overline{T} - \overline{T}') &\cong (1 + m^{-1})\text{tr}(\mathbf{B}\overline{\mathbf{V}}^{-1})/k \end{aligned} \tag{11}$$

Notice that the difference between $\overline{T}$ and $\overline{T}'$ functions much like the $\mathbf{B}$ matrix in the Wald test, capturing between-imputation variation in the estimates (i.e., missing data uncertainty); $\overline{T}$ and $\overline{T}'$ are identical with no missing data, and the difference becomes increasingly positive as the missing data rates increase.

Meng and Rubin's (1992) test statistic differs from its complete-data counterpart in at least two important ways. For one, the procedure will occasionally yield negative values (Liu & Enders, 2016; Schafer, 1997). The $\overline{T}'$ term is usually responsible for this anomalous result, as evaluating the log likelihood at the pooled estimates can produce situations where the restrictive model (i.e., the researcher's hypothesized model) fits better than the general model (i.e., the saturated model). Schafer (1997) suggests that negative values usually occur when the fraction of missing information (i.e., missing data uncertainty, between-imputation variation) is large and the models produce similar fit.[3] This same problem can also cause the difference between $\overline{T}$ and $\overline{T}'$ (and thus $\bar{r}_T$) to take on negative values. Such a result is nonsensical in light of the equivalencies in Equation 11 because a negative value for $\bar{r}_T$ suggests that the estimates have negative between-imputation variance. If the average relative increase in variance is positive and the solution is otherwise admissible (e.g., no negative variance estimates), it may be reasonable to set a negative $T_{(\text{imp})}$ value to zero. However, the test should not be interpreted if $\bar{r}_T$ is negative.

In the context of SEM, it is widely known that the likelihood ratio test is invariant to changes in model parameterization or identification, whereas the Wald test is not (Gonzalez & Griffin, 2001; Satorra, 1989). For example, setting the scale of a latent variable by fixing one of its loadings to unity can produce different Wald z tests than an equivalent identification strategy that constrains the latent variable's variance to one. Because it is scale invariant, the likelihood ratio statistic is unaffected by such reparameterizations. Interestingly, Meng and Rubin's (1992) likelihood ratio test does not possess the scale invariance property of its complete-data counterpart. Schafer (1997) points out that pooling estimates can lead to different results under nonlinear transformations of the parameters. Such transformations occur, for example, when implementing the alternate identification strategies described above (Gonzalez & Griffin, 2001) because certain elements of the model-implied covariance matrix are nonlinear functions of the model parameters (e.g., variances are a function of squared loadings). As such, pooled estimates from equivalent parameterizations will generally produce different estimates of the model-implied covariance matrix and thus different values of $\overline{T}'$ in Equation 8. From a practical perspective, a lack of scale invariance implies that seemingly innocuous identification constraints could impact model fit, such that two equivalent parameterizations of the same model can yield different $T_{(\text{imp})}$ values. Whether alternate parameterizations are capable of introducing meaningful differences in fit is an open question.

Simulation Study

This section describes a Monte Carlo simulation that examines the performance of Meng and Rubin's (1992) likelihood ratio test. A three-factor model with three indicators per factor served as the population model for the study. The data-generating model had a loading matrix of

[3] Most software programs that analyze multiply imputed data (e.g., Mplus, semTools) report fractions of missing information for model parameters, which are just the ratios of between-imputation variability to the total sampling variation (between- plus within-imputation variance).

$$\boldsymbol{\Lambda}^T = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & .65 & .70 & .75 \\ 0 & 0 & 0 & .65 & .70 & .75 & 0 & 0 & 0 \\ .65 & .70 & .75 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad (12)$$

and a factor covariance matrix equal to

$$\Phi = \begin{bmatrix} 1.0 & .45 & .45 \\ .45 & 1.0 & .45 \\ .45 & .45 & 1.0 \end{bmatrix}. \quad (13)$$

The residual covariance matrix was a diagonal matrix with values set to produce manifest variables with unit variances. This model structure is identical to that of Lee and Cai (2012), and the parameter values are similar in magnitude. Importantly, all manifest variables were normally distributed and the analysis models were estimated with normal-theory maximum likelihood. Because Meng and Rubin's (1992) likelihood ratio test assumes a quadratic form for the likelihood, we are applying the test to a context for which it is theoretically well suited.

We employed a simulation design that consisted of three between-subjects factors and one within-subjects factor. The between-subjects factors were missing data mechanism (MCAR and MAR), missing data rate (10%, 20%, 30%, and 40%), and sample size (100, 200, 400, 800, and 1,600), and the number of imputations ($m = 20$ and 100) was the sole within-subjects factor. We chose missing data rates and sample sizes to cover a wide range of values. Given the lack of existing research on this topic, our goal for the simulation was to provide a preliminary investigation of the likelihood ratio test in the context of a relatively simple model. Future studies will certainly extend our understanding of the imputation-based test statistic to a wider range of models and conditions.

We used the R software program to generate 1,000 artificial data sets within each of the 40 between-subjects design cells. Data generation employed the following steps: (a) create an $N \times 9$ matrix of random normal variates; (b) apply Cholesky decomposition to the population variance-covariance matrix and use the resulting matrix to transform the raw data to the desired covariance structure (all means were fixed at zero); and (c) impose a specified proportion of MCAR or MAR missing values. The final step created missing values on all the indicators of the second and third latent variables. As described below, the indicators of the first latent variable served as the causes of missingness in the MAR simulation.

Each indicator of the first factor served as the cause of missingness for two other indicators; the first indicator of the first factor caused missingness on the first indicator of the other two factors, the second indicator caused missingness on the other middle indicators, and so on. The deletion process worked as follows. Using the latent variable formulation for logistic regression (Agresti, 2012; Johnson & Albert, 1999), we derived intercept and slope coefficients from the regression of a missing data indicator on a standardized predictor (the cause of missingness). In the MCAR condition, the intercept determined missingness and the slope equaled zero, and in the MAR condition the slope took on a positive value, such that the probability of missing data increased with values of the first factor's indicators. To ensure a sufficiently strong selection mechanism (i.e., one that would introduce bias if the first factor was omitted from missing data handling), we chose a slope coefficient that produced a squared correlation of .40 between the cause of missingness and the underlying latent propensity for missing data.

After obtaining the logistic coefficients, we substituted the first factor's indicator scores into the regression equation and used the logit link to obtain an $N \times 3$ vector of predicted probabilities, one for each indicator. Finally, we used a binomial distribution function to generate a six-element vector of missing data indicators for each case, where the corresponding predicted probabilities defined the distribution's success rate (i.e., the probability of missing data). Each of the incomplete variables was coding as missing if its corresponding indicator equaled one.

To evaluate the convergence behavior of the Markov chain Monte Carlo (MCMC) algorithm, we examined trace plots and potential scale reduction factors (Gelman & Rubin, 1992) from several samples. We used the information from these diagnostic runs to specify an MCMC imputation algorithm (i.e., data augmentation; Schafer, 1997) in Mplus 7.3 that generated $m = 100$ imputations, saving a data set after every 500th computational cycle. Although recent recommendations from the literature (Graham et al., 2007) suggest that $m = 20$ imputations are sufficient for many analyses, it is well known that quantities such as the fraction of missing information (an algebraic transformation of the relative increase in variance) are unstable and require a much larger number of imputations (Bodner, 2008; Schafer, 1997). Meng and Rubin's (1992) likelihood ratio test does not explicitly incorporate fractions of missing information, but the equivalencies in Equation 11 suggest that the test relies on these quantities, at least indirectly. As such, we generated and analyzed $m = 100$ imputations per replication, and we subsequently examined model fit using a reduced set of $m = 20$ data sets (the first 20 of 100).

Finally, a brief discussion of the imputation model is warranted. An important principle of imputation is that the imputation model should be at least as general (i.e., include at least as many variables or effects, perhaps more) as the analysis model (Meng, 1994; Schafer, 2003). This implies that, in order for Meng and Rubin's (1992) statistic to function as a test of global model fit, the imputation model must preserve associations in both the saturated and hypothesized models. Thus, the imputation model for the simulation was saturated and included the nine indicators from the analysis model. We highlight this issue because using a rich imputation model to treat missing data in a structured model analysis could decrease power relative to FIML, which treats missing data during estimation of the simpler structured model. A similar loss of efficiency has been observed with two-stage maximum likelihood estimation, a procedure that also uses a saturated missing data model (Savalei & Bentler, 2009).

Given the lack of existing research, it is important to study the performance of Meng and Rubin's (1992) test when the fitted model is correct or misspecified. The two panels of Figure 1 show the primary analysis models from the simulation; the top panel is the correct model (i.e., the data-generating model), and the bottom panel incorrectly sets the structural path from *X* to *Y* at zero. Recall that the imputation-based likelihood ratio test is not scale invariant, meaning that its value can potentially vary across different parameterizations of the same model. To examine this possibility, we additionally fit the correctly specified model as a confirmatory factor analysis (CFA; i.e., the model from the top panel of Figure 1 with straight arrows changed to curved arrows). To facilitate the interpretation of the simulation results, we compared Meng and

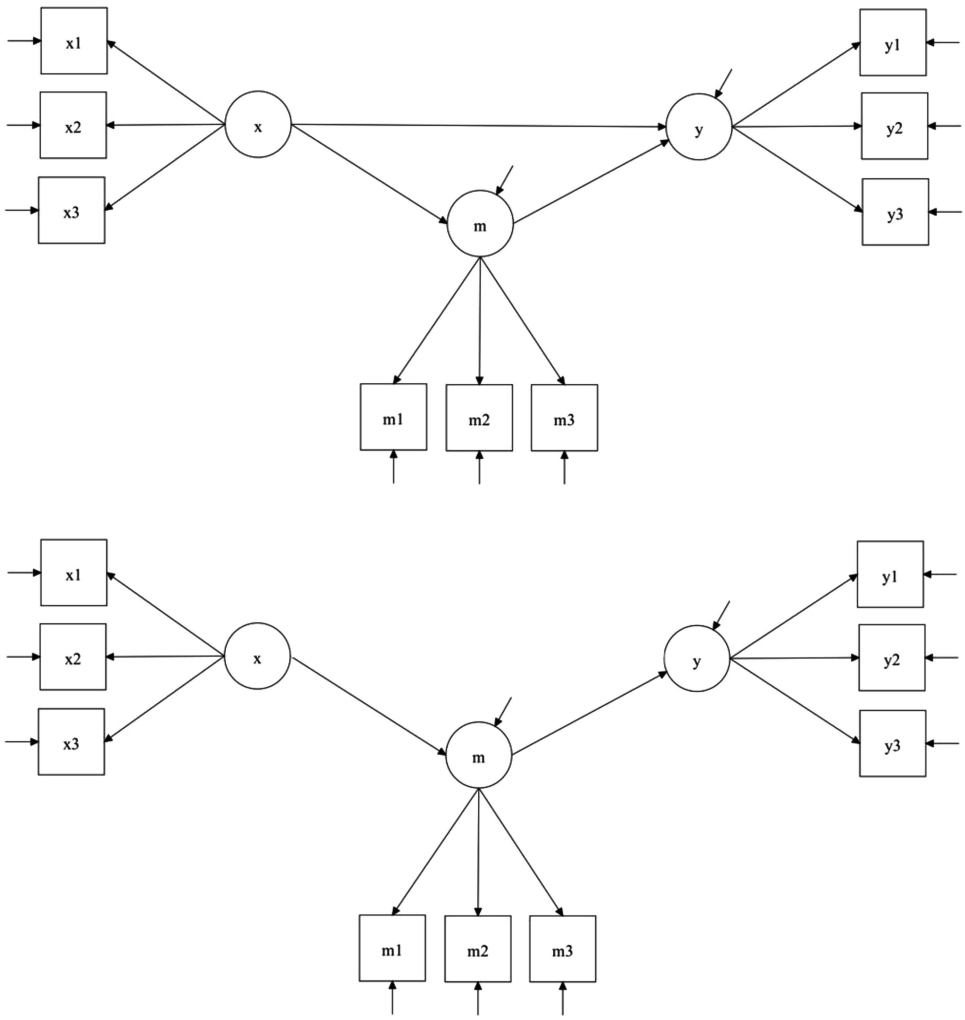

Figure 1. The top model is the population data-generating model and thus yields perfect fit. The bottom panel depicts the misspecified model.

Rubin's (1992) test statistic with that of complete-data and missing-data FIML estimation. As noted previously, FIML test statistics have a large body of supporting research and thus provide a useful baseline for comparison.

We used maximum likelihood estimation in Mplus 7.3 (Muthén & Muthén, 1998–2012) to fit the various analysis models, and we wrote a custom R program that automated the process of estimating models and culling the appropriate output. In small samples, our preliminary test simulations showed that including solutions with negative variance estimates occasionally produced unstable estimates of the average relative increase in variance and the pooled likelihood ratio test statistic (e.g., implausible negative values). Because Mplus currently includes all converged imputations in the computations, we instead saved the parameter estimates and wrote a custom R program to compute the likelihood ratio test after screening out solutions with negative variance estimates and/or replications that produced a negative average relative increase in variance. We applied a similar procedure to FIML, working under the assumption that researchers would not interpret inadmissible solutions in practice. Finally, we omitted from the study replications where fewer than half of the desired imputations converged with admissible estimates. We chose this somewhat arbitrary cutoff because we suspect that many researchers would be wary of interpreting a solution that failed to converge in the majority of data sets. Across all design cells, our exclusion criteria removed about 1% of replications, with most of the problematic cases occurring in the $N = 100$ condition. Perhaps not surprisingly, the percentage of inadmissible solutions increased as a function of missing data, ranging from approximately 1% (10% missing data) to 10% (40% missing data). Inadmissible solutions were rare in the $N = 200$ condition, even with 40% missing data. All simulation scripts are available upon request, as are tabled results from the preliminary simulations that included inadmissible solutions.

We summarized the simulation results by (a) describing the distribution of the likelihood ratio test statistic (e.g., mean, variance, minimum and maximum values) for each analysis method; (b) computing Type I error rates for the correctly specified analysis

model; and (c) examining empirical power to reject the misspecified analysis model. The proportion of significant likelihood ratio tests from each design cell defined the Type I error rate for the true model and empirical power to reject the misspecified model, respectively.

Simulation Results

As explained previously, Meng and Rubin's (1992) likelihood ratio statistic is not scale invariant, meaning that its value can potentially vary across equivalent parameterizations of the same model. This is an important issue for SEM analyses because the test statistic would have limited utility if model fit were impacted by arbitrary identification constraints. To examine the impact of scaling decisions, we fit the structural model from the top panel of Figure 1 and a corresponding CFA (i.e., the model from the top panel of Figure 1 with straight arrows changed to curved arrows). Table 1 gives the mean and variance of the model fit statistics for the MAR simulation broken down sample size, missing data rate, and number of imputations. Because the MCAR and MAR mechanisms were essentially identical, we limit the discussion to the MAR conditions. Full tabular results are available upon request.

As seen in the table, differences between the two parameterizations were generally negligible except for the $N = 100$ design cells with 40% missing data, where the CFA model produced test statistics with slightly smaller variance. Taken as a whole, the simulation results suggest that Meng and Rubin's (1992) test statistic is largely immune to arbitrary scaling decisions, although the choice of parameterization could have a modest impact on inference in small samples with high missing data rates. Given the similarity of the parameterizations, we focus on the structural model from Figure 1 for the remainder of the article.

Performance With a Correctly Specified Analysis Model

We now examine the performance of the imputation-based likelihood ratio test with the correctly specified model shown in the top panel of Figure 1. We use complete-data and FIML test statistics as a benchmark because they have been the focus of a great deal of methodological research. The central chi-square distribution for this particular model has an expected value and variance of 24 and 48, respectively, so these values are also useful for evaluating the results. As before, we limit the discussion to the MAR mechanism because the MCAR results were comparable.

Table 2 gives the mean and variance of the likelihood ratio tests broken down by sample size, missing data rate, and number of imputations. As seen in the tables, missing-data FIML and multiple imputation generally produced similar test statistics, and both procedures were typically well calibrated to the complete data. The largest differences occurred with $N = 100$ and 40% missing rate, where Meng and Rubin's (1992) test statistic exhibited noticeably larger variance when computed from $m = 20$ imputations. An increase in variance was also evident at $N = 200$ and 40% missing data, but this effect was slight compared with $N = 100$. With the exception of the $N = 100$ and 40% missing data rate condition, using fewer imputations had little bearing on the distribution of the multiple imputation test statistic.

To further characterize differences among the test statistics, Table 3 gives the Type I error rates for each design cell of the MAR simulation. In addition to the chi-square metric used by SEM software, we also characterize Type I error rates from the approximate F reference distribution proposed by Li et al. (1991). With 1,000 replications per cell, a .05 proportion would have a standard error of approximately .007, such that the upper and lower limits of .064 and .036, respectively, define a 95% interval around the nominal Type I error rate. As seen in the table, the multiple

Table 1
Comparison of Likelihood Ratio Test Statistics for Two Equivalent Parameterizations of the Correctly Specified Analysis Model (MAR)

		$m = 20$				$m = 100$			
		SEM		CFA		SEM		CFA	
N	Missing	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.
100	10%	25.33	51.60	25.32	51.54	25.33	51.15	25.32	51.10
	20%	25.76	59.57	25.70	59.12	25.77	58.07	25.71	57.62
	30%	25.61	58.34	25.51	60.05	25.53	56.10	25.40	56.87
	40%	26.60	80.52	26.23	74.35	25.82	61.26	25.42	56.41
200	10%	24.74	51.46	24.73	51.45	24.73	50.59	24.73	50.58
	20%	25.18	55.71	25.16	55.62	25.18	55.26	25.16	55.16
	30%	25.00	55.22	24.95	55.00	24.92	53.36	24.87	53.14
	40%	25.28	56.85	25.15	55.50	25.17	52.55	25.04	51.97
400	10%	24.38	51.15	24.38	51.14	24.35	50.96	24.34	50.96
	20%	24.61	53.62	24.60	53.60	24.58	52.47	24.57	52.45
	30%	24.70	53.90	24.68	53.83	24.72	53.20	24.70	53.15
	40%	24.60	53.59	24.56	53.44	24.47	53.35	24.43	53.25
800	10%	24.35	51.98	24.35	51.98	24.30	51.36	24.30	51.36
	20%	23.89	50.13	23.89	50.12	23.84	49.43	23.84	49.42
	30%	24.68	57.60	24.67	57.58	24.64	56.32	24.63	56.30
	40%	24.55	54.04	24.54	54.00	24.50	53.21	24.49	53.17

Table 2
Likelihood Ratio Test Statistics From a Correctly Specified Analysis Model (MAR)

		Complete		FIML		MI (m = 20)		MI (m = 100)	
N	Missing	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.
100	10%	25.39	52.44	25.47	51.91	25.33	51.60	25.33	51.15
	20%	25.27	60.94	26.01	58.55	25.76	59.57	25.77	58.07
	30%	24.82	49.35	26.00	54.54	25.61	58.34	25.53	56.10
	40%	25.34	52.34	27.62	63.24	26.60	80.52	25.82	61.26
200	10%	24.82	52.23	24.83	50.55	24.74	51.46	24.73	50.59
	20%	24.68	53.47	25.37	56.21	25.18	55.71	25.18	55.26
	30%	24.42	48.95	25.06	50.77	25.00	55.22	24.92	53.36
	40%	24.68	43.74	25.32	48.37	25.28	56.85	25.17	52.55
400	10%	24.23	49.96	24.34	50.17	24.38	51.15	24.35	50.96
	20%	24.51	53.65	24.59	52.19	24.61	53.62	24.58	52.47
	30%	24.55	47.00	24.74	51.14	24.70	53.90	24.72	53.20
	40%	24.44	52.13	24.74	52.12	24.60	53.59	24.47	53.35
800	10%	24.14	49.61	24.35	51.48	24.35	51.98	24.30	51.36
	20%	24.03	47.54	23.92	49.05	23.89	50.13	23.84	49.43
	30%	24.79	54.08	24.67	53.99	24.68	57.60	24.64	56.32
	40%	24.18	47.38	24.54	49.36	24.55	54.04	24.50	53.21

imputation test statistic generally produced Type I error rates that were quite close to those of FIML, and there was virtually no difference between the chi-square and F reference distributions. The similarity of the two distributions follows from the fact that the approximate F distribution has a very large denominator degrees of freedom, making its tail probabilities effectively equivalent to those of the corresponding central chi-square (e.g., with 40% missing data, the average degrees of freedom value was approximately 1,200, and this value exceeded 20,000 with 10% missing data). Finally, consistent with the results from Table 2, increasing the number of imputations from 20 to 100 offered a slight improvement with N = 100 and 40% missing data, but this feature otherwise had a negligible impact on Type I error rates.

Performance With a Misspecified Analysis Model

Next, we considered the misspecified model that fixed the structural path from X to Y to zero (see the bottom panel of Figure 1). Tables 4 summarizes the likelihood ratio statistics from this simulation. The complete-data test statistic is no longer a relevant benchmark because its value is systematically higher due to better power. As seen in the table, Meng and Rubin's (1992) test had a slightly lower mean and variance than FIML in most conditions, which suggests that it also has lower power. Interestingly, these distributional differences increased with the sample size, and so the tests do not appear to have the same asymptotic behavior. Consistent with previous results, increasing the number of imputations from 20 to 100 had a negligible impact that was noticeable only at N = 100 with 40% missing data.

Table 3
Type I Error Rates of the Correctly Specified Analysis Model (MAR)

N	Missing	Comp.	FIML	MI (Chi)	MI (F)
100	10%	.073	.072	.077	.077
	20%	.089	.083	.083	.081
	30%	.058	.086	.080	.078
	40%	.080	.134	.112	.110
200	10%	.064	.064	.067	.066
	20%	.062	.086	.079	.079
	30%	.063	.069	.078	.076
	40%	.053	.071	.077	.074
400	10%	.050	.057	.061	.061
	20%	.067	.052	.060	.059
	30%	.055	.066	.072	.068
	40%	.057	.068	.067	.064
800	10%	.057	.061	.058	.058
	20%	.050	.055	.058	.057
	30%	.060	.067	.071	.070
	40%	.044	.065	.069	.068

To explore the practical impact of the mean differences, Table 5 gives empirical power estimates (i.e., the proportion of significant test statistics) from each design cell. In line with Tables 4, the multiple imputation test statistic exhibited lower power, particularly when the missing data rate was 20% or higher. The power differential observed in Table 5 is not necessarily surprising given that imputation uses a saturated model and FIML integrates missing data handling into the estimation of a simpler structural model with 25 fewer parameters. A similar loss of efficiency has been observed with two-stage maximum likelihood estimation, a procedure that also uses a saturated missing data model (Savalei & Bentler, 2009). However, the fact that power differences were more evident at larger sample sizes suggests that additional factors may be at play. We return to this issue later in the Discussion section.

Fit Indices for Multiply Imputed Data

Several fit indices that currently enjoy widespread use are easily computed from the Wald or likelihood ratio test, including the CFI, TLI, and RMSEA, among others. Although the literature offers virtually no guidance on computing SEM fit indices with multiply imputed data, the previous simulation results suggest that Meng and Rubin's (1992) likelihood ratio statistic can serve as the foundation for imputation-based indices. A simple yet unstudied strategy for constructing fit indices with multiply imputed data is to substitute Meng and Rubin's

Table 4
Likelihood Ratio Test Statistics From a Misspecified Analysis Model (MAR)

		FIML		MI (m = 20)		MI (m = 100)	
N	Missing	Mean	Var.	Mean	Var.	Mean	Var.
100	10%	30.37	68.80	29.86	65.51	29.88	65.08
	20%	30.47	73.95	29.61	72.04	29.60	68.93
	30%	30.03	70.53	28.78	69.63	28.76	67.38
	40%	30.92	71.86	29.05	83.75	28.43	62.61
200	10%	33.58	86.43	32.91	84.06	32.93	82.57
	20%	33.43	89.49	32.21	81.60	32.26	80.37
	30%	32.49	82.16	31.01	76.80	31.02	73.96
	40%	31.55	72.12	29.81	69.08	29.84	68.06
400	10%	41.36	120.70	40.21	112.15	40.26	112.14
	20%	39.64	104.43	37.63	91.61	37.75	89.87
	30%	37.70	101.76	35.19	86.05	35.36	86.32
	40%	36.02	94.21	33.04	78.06	33.07	77.79
800	10%	56.99	187.26	54.64	168.95	54.78	170.39
	20%	53.04	164.56	49.08	136.54	49.30	136.94
	30%	50.07	161.05	45.19	129.00	45.48	128.36
	40%	45.98	128.19	40.53	98.82	40.83	98.69
1,600	10%	87.13	284.97	82.48	252.68	82.86	254.69
	20%	80.74	276.83	73.07	229.23	73.68	226.20
	30%	73.68	252.60	64.07	185.39	64.72	186.25
	40%	66.63	222.99	55.82	157.26	56.47	156.55

(1992) test statistic into standard complete-data fit formulas. After a brief review of the TLI, CFI, and RMSEA, we outline estimators that apply this approach. Although the literature describes a vast array of fit measures (e.g., see West, Taylor, & Wu, 2012), we focus on these particular indices because the methodological literature provides guidelines for interpreting their values (Hu & Bentler, 1995, 1999; West et al., 2012) and because they are common in published applications (McDonald & Ho, 2002).

Comparative or incremental fit indices such as the TLI and CFI address the relative adequacy of two nested models, the first of which is a more general model, the second of which is a more restrictive model that places constraints on certain parameters. In SEM applications, the usual scenario pits the researcher's model (the general model) against a null or independence model that posits no correlation among the variables (the restrictive model).[4] We use T_0 and T_I to denote the likelihood ratio (chi-square) test of model fit for the researcher's hypothesized model and the independence model, respectively, where df_0 and df_I are the corresponding degrees of freedom. In contrast, the RMSEA is an absolute index that estimates population misfit of the researcher's model. Common expressions for these indices are given below. Notice that the TLI uses the ratio of the chi-square to its degrees of freedom (i.e., the expected value of a central chi-square distribution) to quantify misfit, whereas the CFI uses the estimated noncentrality parameter (i.e., the difference between the chi-square and its expected value) to express the misfit of each model. Finally, the RMSEA uses the noncentrality parameter from the researcher's hypothesized model to produce an absolute index that quantifies standardized misfit per degree of freedom. A variety of resources provide additional information on these indices (e.g., see West et al., 2012).

$$\text{TLI} = \frac{T_I/\text{df}_I - T_0/\text{df}_0}{T_I/\text{df}_I - 1}$$
$$\text{CFI} = \frac{\max(T_I - \text{df}_I, 0) - \max(T_0 - \text{df}_0, 0)}{\max(T_I - \text{df}_I, 0)} \quad (14)$$
$$\text{RMSEA} = \sqrt{\frac{\max(T_0 - \text{df}_0, 0)}{\text{df}_0(N-1)}}$$

FIML versions of the fit indices are obtained by substituting the appropriate missing-data likelihood ratio statistic into the expressions from Equation 14. The FIML test statistic incorporates missing data uncertainty via its dependence on the observed missing data pattern, where each case's contribution to the log likelihood depends on the number of observed data points for that case. Equation 9 shows that Meng and Rubin's (1992) pooled test statistic achieves a similar goal via a downward adjustment that depends on the average relative increase in variance, a measure of missing data uncertainty based on the variance of the estimates across imputations. Because the imputation-based test statistic appropriately adjusts for missing data and appears to adequately approximate the desired chi-square distribution in most cases, chi-square-based fit indices can be obtained by substituting Meng and Rubin's (1992) test statistic into the previous expressions; this possibility was previously raised by Lee and Cai (2012). Substituting the imputation-based test statistic and estimates of noncentrality give the following expressions.

[4] A null or independence model may not be the appropriate comparison for all situations. See Widaman and Thompson (2003) for a detailed treatment of this topic.

Table 5
Empirical Power From a Misspecified Analysis Model (MAR)

			$m = 20$		$m = 100$	
N	Missing	FIML	MI (Chi)	MI (F)	MI (Chi)	MI (F)
100	10%	.193	.160	.160	.168	.167
	20%	.184	.159	.154	.150	.150
	30%	.177	.127	.126	.138	.138
	40%	.207	.137	.133	.126	.126
200	10%	.292	.269	.268	.272	.272
	20%	.308	.262	.261	.259	.258
	30%	.268	.204	.203	.212	.212
	40%	.227	.161	.159	.160	.160
400	10%	.616	.579	.579	.582	.582
	20%	.547	.462	.462	.473	.473
	30%	.474	.373	.370	.382	.382
	40%	.397	.280	.276	.274	.273
800	10%	.933	.917	.917	.914	.914
	20%	.889	.849	.847	.847	.847
	30%	.839	.725	.724	.734	.734
	40%	.766	.585	.577	.614	.613
1,600	10%	1.000	1.000	1.000	1.000	1.000
	20%	1.000	.992	.992	.993	.993
	30%	.996	.987	.987	.985	.985
	40%	.988	.942	.939	.948	.948

$$\begin{aligned} \text{TLI} &= \frac{\max(T_{\text{I(imp)}}, 0)/\text{df}_{\text{I}} - \max(T_{0(\text{imp})}, 0)/\text{df}_0}{\max(T_{\text{I(imp)}}, 0)/\text{df}_{\text{I}} - 1} \\ \text{CFI} &= \frac{\max(T_{\text{I(imp)}} - \text{df}_{\text{I}}, 0) - \max(T_{0(\text{imp})} - \text{df}_0, 0)}{\max(T_{\text{I(imp)}} - \text{df}_{\text{I}}, 0)} \\ \text{RMSEA} &= \sqrt{\frac{\max(T_{0(\text{imp})} - \text{df}_0, 0)}{\text{df}_0(N-1)}} \end{aligned} \tag{15}$$

The above expressions are identical to those in Equation 14 with one exception: the possibility of a negative test statistic requires a slight modification to the TLI, where $\max(T_{\text{I(imp)}}, 0)$ and $\max(T_{0(\text{imp})}, 0)$ replace T_{I} and T_0, respectively.

Because it lacks scale invariance, Meng and Rubin's (1992) test more closely resembles a Wald statistic than a likelihood ratio test; as explained previously, the authors' derivations could be viewed as a mechanism for estimating a Wald test from m likelihood ratio values. Given its Wald-like properties, the well-known asymptotic equivalence of the Wald and likelihood ratio tests and their corresponding noncentrality parameters (Satorra, 1989) provides a rationale for the expressions in Equation 15, with the caveat that such equivalence holds only when a model is correctly specified (Yuan & Chan, 2005). It is also worth noting that Bentler (1990) outlined Wald-based versions of the TLI and CFI. This work provides an additional rationale for the above indices, again with the caveat that different estimates of noncentrality can yield different estimates of fit statistics (Yuan & Chan, 2005).

Simulation Results: Fit Indices

To examine the behavior of imputation-based fit indices, we applied the expressions from Equation 15 to the artificial data sets from the previous computer simulation. Although the SRMR has no relation to the likelihood ratio test, we include this index for completeness. For brevity, we again focus on the MAR results because the mechanism had no material impact on the findings. Finally, we focus on the $m = 20$ condition for the remainder of the article because the number of imputations had a negligible impact on the likelihood ratio test, and 20 imputations is consistent with current recommendations from the literature (Graham et al., 2007). Complete tabular results are available upon request.

To begin, consider the fit indices for the properly specified analysis model. Tables 6 through 9 give the CFI, TLI, RMSEA, and SRMR values, respectively, broken down by sample size and missing data rate. As seen in Table 9, mean CFI values for FIML and multiple imputation were slightly lower than those of the complete data at sample sizes of 200 or lower and missing data rates of 30% or higher. The imputation-based CFI was particularly low in the $N = 100$ and 40% missing data condition, which is not surprising in light of the previous results. Table 7 shows a similar pattern of results for the TLI, except that differences were primarily evident at $N = 100$. Table 8 shows that the missing-data RMSEA values were well calibrated to the complete data, and differences between FIML and imputation were generally nil. The $N = 100$ conditions with 30% or higher missing data rates introduced a slight positive bias in the missing-data RMSEAs, but this tendency disappeared at larger sample sizes. Finally, Table 9 shows that FIML and imputation-based SRMR values overestimated those of the complete data (even at a sample size of $N = 800$), but the two procedures gave virtually identical estimates in all conditions. This result is encouraging in light of the Yuan, Yang-Wallentin, and Bentler (2012) study that reported bias in imputation-based estimates of covariances.

Tables 10, 11, 12, and 13 summarize the fit indices for the misspecified analysis model. For the CFI and TLI, results were similar to those of the correctly specified model. Specifically, the missing-data estimates were generally indistinguishable from the complete-data estimates (and from each other), except at $N = 100$ with 30% or higher missing data. Owing to a decrease in power, the FIML and imputation-based estimates of the RMSEA were slightly lower than those of the complete data. The pattern of

Table 6
CFI Summary for the Correctly Specified Analysis Model (MAR)

		Complete		FIML		MI (m = 20)	
N	Missing	Mean	*SD*	Mean	*SD*	Mean	*SD*
100	10%	.985	.022	.983	.025	.982	.027
	20%	.985	.023	.979	.030	.975	.035
	30%	.986	.020	.976	.031	.969	.045
	40%	.985	.021	.966	.041	.949	.086
200	10%	.993	.010	.992	.012	.992	.012
	20%	.993	.011	.990	.015	.989	.017
	30%	.994	.010	.990	.015	.987	.021
	40%	.994	.009	.988	.017	.982	.029
400	10%	.997	.005	.996	.006	.996	.006
	20%	.997	.005	.996	.007	.995	.008
	30%	.997	.005	.995	.008	.994	.010
	40%	.997	.005	.994	.009	.992	.012
800	10%	.999	.002	.998	.003	.998	.003
	20%	.999	.002	.998	.003	.998	.003
	30%	.998	.003	.998	.004	.997	.005
	40%	.999	.002	.997	.004	.996	.006

means from Table 4 carried over to the RMSEA, such that imputation-based RMSEA values were somewhat lower than those of FIML with large samples and high missing data rates. Finally, FIML and imputation-based SRMR values overestimated those of the complete data at sample sizes of 200 or lower, but the estimates were otherwise well calibrated; again, the missing data handling methods were effectively equivalent with respect to the SRMR.

Analysis Example

To illustrate the application of Meng and Rubin's (1992) test, we fit the model in the top panel of Figure 1 to one of the artificial data sets from the simulation study (N = 400 with 20% missing data). Appendices A and B give the Mplus syntax files for the imputation and analysis phases, respectively, and Appendix C gives the corresponding semTools syntax for R. The semTools code relies on the MICE package for imputation (van Buuren & Groothuis-Oudshoorn, 2011) and the lavaan package (Rosseel, 2012) for the analysis. The syntax files and raw data for the example are also available for download at www.appliedmissingdata.com.

As noted previously, our test simulations indicated that the likelihood ratio statistic is quite sensitive to solutions producing negative variance estimates, and so these data sets should be excluded from analysis. The TECH9 option of the Mplus OUTPUT command is important for this purpose because it prints imputation-specific warnings and errors, some of which are typically triggered by inadmissible solutions. If such errors occur, the corresponding imputations can be excluded by simply editing the listing file that serves as input data for the analysis. As of this writing, the current version of semTools (version 0.5–12) excludes all improper solutions when computing the test, and the package offers an argument that gives the user explicit control over the

Table 7
TLI Summary for the Correctly Specified Analysis Model (MAR)

		Complete		FIML		MI (m = 20)	
N	Missing	Mean	*SD*	Mean	*SD*	Mean	*SD*
100	10%	.993	.048	.991	.055	.992	.058
	20%	.994	.051	.986	.064	.986	.077
	30%	.996	.046	.984	.069	.983	.099
	40%	.993	.047	.966	.081	.962	.163
200	10%	.998	.023	.998	.026	.998	.028
	20%	.998	.024	.995	.031	.995	.036
	30%	.999	.023	.996	.034	.995	.045
	40%	.998	.022	.994	.038	.992	.061
400	10%	1.000	.012	.999	.013	.999	.014
	20%	.999	.012	.999	.015	.999	.017
	30%	.999	.011	.998	.017	.998	.022
	40%	.999	.012	.998	.019	.998	.028
800	10%	1.000	.006	1.000	.007	1.000	.007
	20%	1.000	.006	1.000	.007	1.000	.008
	30%	.999	.006	.999	.009	.999	.011
	40%	1.000	.006	.999	.009	.999	.014

Table 8
RMSEA Summary for the Correctly Specified Analysis Model (MAR)

		Complete		FIML		MI (m = 20)	
N	Missing	Mean	*SD*	Mean	*SD*	Mean	*SD*
100	10%	.026	.028	.026	.029	.026	.028
	20%	.026	.029	.028	.030	.028	.030
	30%	.024	.028	.028	.029	.027	.030
	40%	.026	.028	.034	.031	.030	.032
200	10%	.017	.020	.017	.020	.017	.020
	20%	.017	.020	.018	.021	.018	.021
	30%	.016	.019	.017	.020	.017	.020
	40%	.016	.019	.018	.020	.018	.021
400	10%	.011	.014	.011	.014	.011	.014
	20%	.012	.014	.012	.014	.012	.014
	30%	.011	.013	.012	.014	.012	.014
	40%	.012	.014	.012	.014	.012	.014
800	10%	.008	.009	.008	.010	.008	.010
	20%	.007	.009	.008	.009	.007	.009
	30%	.009	.010	.008	.010	.009	.010
	40%	.008	.009	.008	.010	.008	.010

inclusion or exclusion of such solutions (S. Pornprasertmanit, personal communication, June 8, 2016).

Because Mplus and semTools generate different imputed data sets, the model fit results will naturally differ; the model fit statistic from Mplus is $\chi^2(24) = 24.698$, $p = .422$, and the corresponding value from semTools is $\chi^2(24) = 24.879$, $p = .412$. Whereas Mplus provides the test statistic on the chi-square metric, semTools also prints the test with an F reference distribution from Li et al. (1991). Substituting the Mplus test statistic into the fit index expressions gives TLI = .998, CFI = .999, and RMSEA = .009.

Discussion

A good deal of recent research has focused on FIML estimation and inference, and the properties of FIML test statistics are generally well understood. Interestingly, very little is known about multiple imputation inference. Much of the existing research has focused on improving the small-sample properties of single-parameter tests (Barnard & Rubin, 1999; Reiter, 2007), and few studies have examined Wald and likelihood ratio tests for multiply imputed data. The application of multiple imputation inference to SEM is an important methodological issue that has received virtually no attention in the literature. This gap in this literature is important because there are a number of practical reasons why researchers might choose multiple imputation over FIML to address missing data in an SEM (e.g., mixtures of categorical and continuous variables, models that use questionnaire items and possibly item parcels). Given the widespread application of SEMs in the behavioral science literature, the goal of this study was twofold: evaluate the application of Meng and Rubin's (1992) pooling procedure for likelihood ratio tests, and explore the possibility of using their test statistic to define imputation-based versions of common fit indices such as the TLI, CFI, and RMSEA.

Table 9
SRMR Summary for the Correctly Specified Analysis Model (MAR)

		Complete		FIML		MI (m = 20)	
N	Missing	Mean	*SD*	Mean	*SD*	Mean	*SD*
100	10%	.048	.009	.053	.009	.052	.009
	20%	.048	.009	.060	.010	.057	.010
	30%	.048	.008	.067	.012	.062	.011
	40%	.048	.008	.079	.014	.071	.013
200	10%	.034	.006	.037	.006	.037	.006
	20%	.034	.006	.042	.007	.041	.007
	30%	.033	.006	.046	.008	.044	.007
	40%	.034	.005	.053	.009	.050	.008
400	10%	.024	.004	.026	.004	.026	.004
	20%	.024	.004	.029	.005	.028	.005
	30%	.024	.004	.032	.005	.031	.005
	40%	.024	.004	.036	.006	.035	.006
800	10%	.017	.003	.018	.003	.018	.003
	20%	.017	.003	.020	.003	.020	.003
	30%	.017	.003	.023	.004	.022	.004
	40%	.017	.003	.025	.004	.025	.004

Table 10
CFI Summary for the Misspecified Analysis Model (MAR)

		Complete		FIML		MI (m = 20)	
N	Missing	Mean	*SD*	Mean	*SD*	Mean	*SD*
100	10%	.971	.030	.969	.033	.969	.034
	20%	.971	.031	.965	.038	.962	.044
	30%	.972	.030	.962	.041	.956	.055
	40%	.972	.029	.954	.046	.938	.095
200	10%	.978	.018	.978	.020	.978	.021
	20%	.978	.019	.975	.023	.975	.025
	30%	.978	.018	.975	.025	.973	.029
	40%	.979	.017	.974	.025	.969	.035
400	10%	.981	.012	.980	.013	.980	.013
	20%	.981	.011	.980	.013	.980	.014
	30%	.981	.012	.980	.015	.979	.017
	40%	.981	.012	.980	.016	.978	.020
800	10%	.981	.007	.980	.008	.981	.008
	20%	.981	.008	.981	.009	.981	.009
	30%	.981	.008	.980	.010	.981	.011
	40%	.981	.007	.981	.010	.981	.012

We focus on Meng and Rubin's (1992) procedure because it is currently available in software packages (Mplus and the semTools package in R) and thus provides researchers with a readily available tool for evaluating model fit. The application of the pooled likelihood ratio test to SEM introduces unique considerations that are not germane to FIML estimation. For example, Meng and Rubin's test is not scale invariant, meaning that two equivalent parameterizations of the same model can yield different test statistics. We investigated this issue by fitting alternate parameterizations of the same model (a full structural model and an equivalent CFA) to simulated data, and the impact of scaling decisions were nil. However, the likelihood ratio statistic was quite sensitive to data sets that produce negative variance estimates, particularly at smaller sample sizes. Including such solutions can give implausible negative values for the average relative increase in variance, which in turn affects the validity of the resulting test statistic. Consequently, we strongly encourage researchers to screen the imputation-specific parameter estimates and analyze only those data sets producing admissible solutions.

Considered as a whole, the simulations suggest that the pooled likelihood ratio statistic gives relatively good performance when used to assess the global model fit of an SEM. The consistent exception occurred at $N = 100$ with 40% missing data, where the imputation-based test consistently differed from that of FIML. Meng and Rubin's derivations are asymptotic, and the previous literature offers no guidance about the test's performance in finite samples. We speculate that the performance of the test statistic at $N = 100$ and high missing data rates has to do with the instability of the average relative increase in variance (the denominator of the test statistic), a quantity that captures between-imputation variability in the estimates. Future studies should investigate the relative increase in variance, as this quantity plays a pivotal role in the behavior of the test statistic.

Table 11
TLI Summary for the Misspecified Analysis Model (MAR)

		Complete		FIML		MI (m = 20)	
N	Missing	Mean	*SD*	Mean	*SD*	Mean	*SD*
100	10%	.967	.053	.964	.059	.966	.061
	20%	.967	.055	.959	.068	.960	.080
	30%	.968	.053	.958	.074	.958	.102
	40%	.967	.051	.945	.082	.940	.164
200	10%	.971	.029	.970	.032	.971	.034
	20%	.971	.030	.967	.037	.968	.041
	30%	.971	.029	.967	.041	.967	.050
	40%	.971	.027	.968	.043	.966	.063
400	10%	.972	.018	.971	.019	.972	.020
	20%	.972	.017	.971	.020	.972	.021
	30%	.972	.017	.971	.023	.972	.026
	40%	.972	.017	.972	.025	.971	.032
800	10%	.973	.011	.972	.012	.973	.012
	20%	.973	.011	.972	.013	.973	.013
	30%	.972	.011	.972	.014	.972	.016
	40%	.973	.011	.973	.014	.973	.017

Table 12
RMSEA Summary for the Misspecified Analysis Model (MAR)

		Complete		FIML		MI (m = 20)	
N	Missing	Mean	*SD*	Mean	*SD*	Mean	*SD*
100	10%	.042	.032	.040	.031	.038	.031
	20%	.041	.033	.040	.032	.037	.032
	30%	.040	.032	.039	.032	.034	.031
	40%	.041	.032	.043	.031	.035	.032
200	10%	.039	.023	.036	.023	.035	.023
	20%	.038	.024	.036	.024	.033	.023
	30%	.039	.023	.033	.023	.030	.023
	40%	.039	.022	.031	.023	.027	.022
400	10%	.040	.015	.038	.016	.036	.016
	20%	.040	.014	.036	.015	.033	.015
	30%	.040	.015	.032	.016	.029	.016
	40%	.040	.015	.030	.016	.025	.016
800	10%	.041	.009	.039	.009	.037	.009
	20%	.041	.009	.036	.009	.033	.009
	30%	.041	.009	.034	.010	.030	.010
	40%	.041	.009	.031	.010	.026	.010

If one puts aside the small sample size and high missing data rate condition, the primary weakness of the imputation-based fit statistic was a lack of power relative to FIML. As noted previously, part of the power differential likely stems from the fact that imputation uses a saturated model, whereas FIML integrates missing data handling into the estimation of a simpler structural model with fewer parameters (24 or 25 fewer parameters in the two models examined here). A similar loss of efficiency has been observed with two-stage maximum likelihood estimation, a procedure that also uses a saturated missing data model (Savalei & Bentler, 2009). At least two issues are worth considering. First, it is not strictly necessary to use a saturated model for imputation, and using a structured model (e.g., a CFA) for imputation might improve power. However, employing a restrictive imputation model would limit the range of models that could be estimated and compared (e.g., a saturated model should not be estimated from imputations generated by a CFA model). Second, an oft-stated benefit of multiple imputation is the ease with which it accommodates a potentially large number of additional auxiliary variables. However, the power differences we observed suggest that including too many (or perhaps the wrong) auxiliary variables could be detrimental to power. Future studies should investigate the impact of auxiliary variables on model fit, comparing Meng and Rubin's (1992) test with the saturated correlates approach in the FIML framework (Graham, 2003). Whether the FIML and imputation-based test statistics are differentially impacted by auxiliary variables is an important practical question for substantive research.

Although model complexity is likely an important determinant of power, the fact that the mean and variance of the test statistic decreased at larger sample sizes suggests that additional factors could also be at play. Two aspects of Meng and Rubin's (1992) derivations should be considered. First, the derivations rely on the asymptotic equivalence of the Wald and likelihood ratio tests, but this equivalence no longer holds in misspecified models such as

Table 13
SRMR Summary for the Misspecified Analysis Model (MAR)

		Complete		FIML		MI (m = 20)	
N	Missing	Mean	*SD*	Mean	*SD*	Mean	*SD*
100	10%	.067	.015	.070	.015	.070	.016
	20%	.066	.015	.072	.015	.073	.016
	30%	.066	.015	.079	.016	.079	.017
	40%	.067	.016	.091	.017	.088	.017
200	10%	.056	.012	.057	.012	.058	.012
	20%	.057	.013	.060	.012	.061	.013
	30%	.056	.012	.062	.012	.064	.013
	40%	.056	.013	.067	.012	.070	.014
400	10%	.050	.010	.051	.010	.051	.010
	20%	.050	.010	.051	.010	.052	.010
	30%	.051	.010	.053	.010	.055	.011
	40%	.050	.010	.054	.010	.058	.011
800	10%	.047	.007	.047	.008	.047	.008
	20%	.047	.007	.047	.007	.048	.008
	30%	.047	.008	.047	.008	.050	.008
	40%	.047	.007	.047	.007	.051	.008

that in Figure 1 (Yuan & Chan, 2005). The derivations further assume that the relative increase in variance (i.e., the ratio of between-imputation variance to within-imputation variance) is the same for all parameters, but the missing data pattern we implemented introduced rather larger variation in this ratio (e.g., the first factor's indicators were complete, and thus its loadings were unaffected by missing data, whereas the second factor's loadings had considerable between-imputation variability). Distributing the missing values equally across all indicators would better approximate the assumption, as would modifying model parameters to achieve a more uniform pattern of model-implied correlations. Additional work is needed to understand the practical importance of these assumptions.

Using Meng and Rubin's (1992) test statistic as the basis for computing fit indices such as the CFI, TLI, and RMSEA appears to be a promising approach. Interestingly, imputation-based comparative fit indices were comparable with those of FIML, even in situations where the pooled test statistic lacked power. The downward bias of the chi-square that occurred with misspecified analysis models also reduced the magnitude of the RMSEA, but this effect was salient only at very high missing data rates (30%–40%). Given that the FIML and imputation-based fit indices were well calibrated to those of the complete data, it seems reasonable to tentatively recommend the common cutoff criteria from the complete-data literature (Hu & Bentler, 1995, 1999). However, it is important to note that the simulations examined a misspecification with a relatively small effect size, so further study is needed to evaluate model selection issues.

As with any simulation, this study has a number of limitations, and its results suggest several avenues for future research. For one, the population model that we used, while identical in complexity to that of Lee and Cai (2012), is probably far simpler than most models that researchers employ in practice. Given the lack of previous studies, our goal for the simulation was to provide a preliminary investigation of Meng and Rubin's (1992) test statistic in the context of a relatively simple, well-behaved model. Future studies should expand this work to include additional conditions, different covariance structures, and models of varying complexity, among other things. Along similar lines, the MAR mechanism for the simulation was a simple logistic function of observed variables in the model. Examining the test statistic with more complex MAR functions and a not missing at random (NMAR) mechanism would be an important avenue to pursue. In particular, very little is known about model fit with NMAR mechanisms, so future studies should investigate sensitivity analyses for model selection. Related to a previous point, future research should investigate NMAR mechanisms that result from excluding auxiliary variables that predict missingness, as the decision to include such variables could involve a tradeoff between bias and precision. Paralleling the development of the complete-data and FIML fit literature, future studies should also examine nonnormal and categorical outcomes, as these features could impact the likelihood ratio statistic, perhaps in a different fashion than with FIML estimation. For example, in the context of a simple bivariate model, Yuan et al. (2012) found that nonnormal data differentially impacted FIML and MI, and it is important to understand whether these differences extend to the SEM context. Further, the asymptotic derivation of Meng and Rubin's (1992) test assumes a quadratic form for the likelihood, so it is important to evaluate the statistic with density functions that do not conform to this ideal (e.g., logistic models). Finally, an interesting avenue of future research is to compare Meng and Rubin's (1992) test with the two-stage approach proposed by Lee and Cai (2012).

In sum, the issue of assessing model fit with multiply imputed data is an important topic that has received surprisingly little attention in the methodological literature. This article examined the application of Meng and Rubin's (1992) pooling procedure for likelihood ratio statistics. Overall, the simulation results suggest that the test statistic holds promise for SEM applications, although caution is definitely warranted when applying the procedure to small samples with extreme missing data rates. Future studies are needed to extend our understanding of the imputation-based test statistic in a wider range of models and conditions.

References

Agresti, A. (2012). *Categorical data analysis* (3rd ed.). Hoboken, NJ: Wiley.

Asparouhov, T., & Muthén, B. (2010). *Chi-square statistics with multiple imputation*. Retrieved from https://www.statmodel.com/download/MI7.pdf

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

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238–246. http://dx.doi.org/10.1037/0033-2909.107.2.238

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588–606. http://dx.doi.org/10.1037//0033-2909.88.3.588

Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling, 15,* 651–675. http://dx.doi.org/10.1080/10705510802339072

Bollen, K. A., & Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods & Research, 21,* 205–229. http://dx.doi.org/10.1177/0049124192021002004

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21,* 230–258. http://dx.doi.org/10.1177/0049124192021002005

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). Thousand Oaks, CA: Sage.

Buse, A. (1982). The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note. *The American Statistician, 36,* 153–157.

Carpenter, J. R., & Kenward, M. G. (2013). *Multiple imputation and its application*. West Sussex, UK: Wiley. http://dx.doi.org/10.1002/9781119942283

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6,* 330–351. http://dx.doi.org/10.1037/1082-989X.6.4.330

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Enders, C. K., Mistler, S. A., & Keller, B. T. (2016). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods, 21,* 222–240. http://dx.doi.org/10.1037/met0000063

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7,* 457–472. http://dx.doi.org/10.1214/ss/1177011136

Gonzalez, R., & Griffin, D. (2001). Testing parameters in structural equation modeling: Every "one" matters. *Psychological Methods, 6,* 258–269. http://dx.doi.org/10.1037/1082-989X.6.3.258

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling, 10,* 80–100. http://dx.doi.org/10.1207/S15328007SEM10014

Graham, J. W. (2012). *Missing data: Analysis and design*. New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4614-4018-5

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8,* 206–213. http://dx.doi.org/10.1007/s11121-007-0070-9

Hu, L., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76–99). Thousand Oaks, CA: Sage.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55. http://dx.doi.org/10.1080/10705519909540118

Johnson, V. E., & Albert, J. H. (1999). *Ordinal data modeling*. New York, NY: Springer.

Lee, T., & Cai, L. (2012). Alternative multiple imputation inference for mean and covariance structure modeling. *Journal of Educational and Behavioral Statistics, 37,* 675–702. http://dx.doi.org/10.3102/1076998612458320

Li, K. H., Raghunathan, T. E., & Rubin, D. B. (1991). Large-sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association, 86,* 1065–1073.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Hoboken, NJ: Wiley. http://dx.doi.org/10.1002/9781119013563

Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods, 18,* 285–300. http://dx.doi.org/10.1037/a0033266

Liu, Y., & Enders, C. K. (2016). *Evaluation of multi-parameter test statistics for multiple imputation*. Manuscript submitted for publication.

McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7,* 64–82. http://dx.doi.org/10.1037/1082-989X.7.1.64

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science, 9,* 538–558.

Meng, X.-L., & Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika, 79,* 103–111. http://dx.doi.org/10.1093/biomet/79.1.103

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

Pornprasertmanit, S., Miller, P., Schoemann, A., Rossee, Y., Quick, C., Garnier-Villarreal, M., . . . Longo, Y. (2016). semTools: Useful tools for structural equation modeling. R package version 0.4–9. Retrieved from http://cran.r-project.org/web/packages/semTools/index.html

Reiter, J. P. (2007). Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. *Biometrika, 94,* 502–508. http://dx.doi.org/10.1093/biomet/asm028

Reiter, J. P., & Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association, 102,* 1462–1471. http://dx.doi.org/10.1198/016214507000000932

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48,* 1–36. http://dx.doi.org/10.18637/jss.v048.i02

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63,* 581–592. http://dx.doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1987a). *Multiple Imputation for Nonresponse in Surveys*. Hoboken, NJ: Wiley. http://dx.doi.org/10.1002/9780470316696

Rubin, D. B. (1987b). *Multiple imputation for nonrespose in surveys*. New York, NY: Wiley. http://dx.doi.org/10.1002/9780470316696

Satorra, A. (1989). Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrika, 54,* 131–151. http://dx.doi.org/10.1007/BF02294453

Savalei, V., & Bentler, P. M. (2009). A two-stage approach to missing data: Theory and application to auxiliary variables. *Structural Equation Modeling, 16,* 477–497. http://dx.doi.org/10.1080/10705510903008238

Savalei, V., & Falk, C. F. (2014). Robust two-stage approach outperforms robust full information maximum likelihood with incomplete nonnormal data. *Structural Equation Modeling, 21,* 280–302. http://dx.doi.org/10.1080/10705511.2014.882692

Savalei, V., & Rhemtulla, M. (2014). *Two-stage estimator for models with composites or parcels when data are missing at the item level*. Paper presented at the Society of Multivariate Experimental Psychology, Nashville, TN.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Boca Raton, FL: Chapman & Hall. http://dx.doi.org/10.1201/9781439821862

Schafer, J. L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica, 57,* 19–35. http://dx.doi.org/10.1111/1467-9574.00218

Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods, 6,* 317–329. http://dx.doi.org/10.1037/1082-989X.6.4.317

Steele, R. J., Wang, N., & Raftery, A. E. (2010). Inference from multiple imputation for missing data using mixtures of normals. *Statistical Methodology, 7,* 351–364. http://dx.doi.org/10.1016/j.stamet.2010.01.003

Steiger, J. H. (1989). EZPATH: A supplementary module for SYSTAT and SYGRAPH. Evanston, IL: SYSTAT.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25,* 173–180.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically-based tests for the number of common factors*. Annual Meeting of the Psychometric Society, Iowa City, IA.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1–10. http://dx.doi.org/10.1007/BF02291170

van Buuren, S. (2012). *Flexible imputation of missing data*. New York, NY: Chapman & Hall. http://dx.doi.org/10.1201/b11826

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45,* 1–67.

von Hippel, P. T. (2015). New confidence intervals and bias comparisons show that maximum likelihood can beat multiple imputation in small samples. *Structural Equation Modeling: A Multidisciplinary Journal, 23,* 422–437. http://dx.doi.org/10.1080/10705511.2015.1047931

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209–231). New York, NY: Guilford Press.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine, 30,* 377–399. http://dx.doi.org/10.1002/sim.4067

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. http://dx.doi.org/10.1037/1082-989X.8.1.16

Yuan, K.-H. (2009). Normal distribution based pseudo ML for missing data: With applications to mean and covariance structure analysis. *Journal of Multivariate Analysis, 100,* 1900–1918. http://dx.doi.org/10.1016/j.jmva.2009.05.001

Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis for nonnormal missing data. *Sociological Methodology, 30,* 165–200. http://dx.doi.org/10.1111/0081-1750.00078

Yuan, K. H., & Bentler, P. M. (2010). Consistency of normal distribution based pseudo maximum likelihood estimates when data are missing at random. *The American Statistician, 64,* 263–267. http://dx.doi.org/10.1198/tast.2010.09203

Yuan, K.-H., & Chan, W. (2005). On nonequivalence of several procedures of structural equation modeling. *Psychometrika, 70,* 791–798. http://dx.doi.org/10.1007/s11336-001-0930-9

Yuan, K.-H., & Savalei, V. (2014). Consistency, bias and efficiency of the normal-distribution-based MLE: The role of auxiliary variables. *Journal of Multivariate Analysis, 124,* 353–370. http://dx.doi.org/10.1016/j.jmva.2013.11.006

Yuan, K.-H., Tong, X., & Zhang, Z. (2014). Bias and efficiency for SEM with missing data and auxiliary variables: Two-stage robust method versus two-stage ML. *Structural Equation Modeling, 22,* 178–192. http://dx.doi.org/10.1080/10705511.2014.935750

Yuan, K.-H., Yang-Wallentin, F., & Bentler, P. M. (2012). ML versus MI for missing data with violation of distributional conditions. *Sociological Methods & Research, 41,* 598–629. http://dx.doi.org/10.1177/0049124112460373

Yuan, K.-H., & Zhang, Z. (2012). Robust structural equation modeling with missing data and auxiliary variables. *Psychometrika, 77,* 803–826. http://dx.doi.org/10.1007/s11336-012-9282-4

Appendix A

Mplus Syntax for Imputation Phase

```
data:
file = example.dat;
variable:
names = x1-x3 m1-m3 y1-y3;
usevariables = x1-y3;
missing = all(999);
analysis:
type = basic;
bseed = 90291;
fbiterations = 500;
data imputation:
impute = m1-y3
ndatasets = 100;
save = imp*.dat;
thin = 500;
output:
tech8;
```

Appendix B

Mplus Syntax for Analysis Phase

```
data:
file = implist.dat;
type = imputation;
variable:
names = x1-x3 m1-m3 y1-y3;
usevariables = x1-y3;
model:
x by x1-x3;
m by m1-m3;
y by y1-y3;
m on x;
y on x m;
output:
tech9;
```

(*Appendices continue*)

Appendix C

semTools Syntax for Imputation and Analysis Phases

```
# load packages
library(mice)
library(lavaan)
library(Contributors)
# import data and name variables
example <- read.table("~/desktop/analysis/example.dat", quote="\"")
example[example == 999] <- NA
names(example) = c('x1','x2','x3','m1','m2','m3','y1','y2','y3')
# lavaan model syntax
model <- '
# measurement model
x =~ x1 + x2 + x3
m =~ m1 + m2 + m3
y =~ y1 + y2 + y3
# regressions
m ~ x
y ~ x + m
'
fim1 <- sem(model, data = example, mimic = "mplus")
summary(fim1, standardized = T)
# impute data, fit model, summarize results
imp <- sem.mi(model, data = example, m = 100, miArgs = list(maxit = 500, seed = 90095),
  miPackage = 'mice', chi = 'all')
summary(imp, standardized = T)
inspect(imp, "fit")
inspect(imp, "impute")
```

Received July 24, 2015
Revision received June 15, 2016
Accepted June 20, 2016 ■