Journal of Educational and Behavioral Statistics
Vol. XX, No. X, pp. 1–27
DOI: 10.3102/1076998617694880

Normal Theory Two-Stage ML Estimator When Data Are Missing at the Item Level

Victoria Savalei
University of British Columbia

Mijke Rhemtulla
University of California, Davis

In many modeling contexts, the variables in the model are linear composites of the raw items measured for each participant; for instance, regression and path analysis models rely on scale scores, and structural equation models often use parcels as indicators of latent constructs. Currently, no analytic estimation method exists to appropriately handle missing data at the item level. Item-level multiple imputation (MI), however, can handle such missing data straightforwardly. In this article, we develop an analytic approach for dealing with item-level missing data—that is, one that obtains a unique set of parameter estimates directly from the incomplete data set and does not require imputations. The proposed approach is a variant of the two-stage maximum likelihood (TSML) methodology, and it is the analytic equivalent of item-level MI. We compare the new TSML approach to three existing alternatives for handling item-level missing data: scale-level full information maximum likelihood, available-case maximum likelihood, and item-level MI. We find that the TSML approach is the best analytic approach, and its performance is similar to item-level MI. We recommend its implementation in popular software and its further study.

Keywords: *item-level missing data; structural equation modeling; two-stage estimation; multiple imputation*

Missing data are common in behavioral research, particularly in studies that take place over time, take place outside of the lab, or involve collecting a lot of data from participants. Frequently, the model of interest is one that posits linear relationships among the variables, such as a regression model, a path analysis model, or a structural equation model (SEM) with latent variables. It is also a common occurrence that the model of interest is at the level of *composites*, that is, linear combinations of the original variables on which data are collected. For example, a regression or a path analysis model may involve predicting one scale score from several other scale scores, where each scale score is computed as a sum of the individual scale items. Another example is an SEM that uses *parcels*,

or sums of several raw items, as indicators of latent factors. Parcels are often recommended when the sample size is small or the number of indicators per factor is large (e.g., Little, Rhemtulla, Gibson, & Schoemann, 2013). In this article, we address the problem of missing data when data are gathered at the item level but are analyzed at the composite level.

We will assume ignorable missing data at the item level (Little & Rubin, 2002), which means that the probability that an observation is missing does not depend on the missing values themselves, conditioning on other variables in the data set. However, even when data are ignorable at the item level, using a suboptimal method to create the composites from the components can lead to nonignorable missingness, as discussed later. Software programs that are capable of fitting SEMs provide two modern ways of dealing with ignorable missing data: full information maximum likelihood (FIML) estimation (Allison, 2003; Arbuckle, 1996; Little & Rubin, 2002) and multiple imputation (MI; Rubin, 1987; Schafer, 1997). Studies comparing FIML and MI in the context of SEM have found the approaches to be largely equivalent, when the number of imputations is large (Collins, Schafer, & Kam, 2001; Larsen, 2011; Lawrence & Lee, 2014; Yuan, Yang-Wallentin, & Bentler, 2012). The choice between FIML and MI is thus often a matter of convenience and availability in software. Arguably, however, SEMs are more easily estimated using FIML. Under the conditions of multivariate normality, the FIML estimator is also asymptotically efficient, while MI has this property only if the number of imputations is infinite. However, when data are gathered at the item level but are analyzed at the composite level, the FIML estimator is no longer available because the variables containing missingness are not directly in the model.[1] Yet, MI at the item level followed by model fitting to composites is straightforward. Thus, MI appears to have an advantage in the case of item-level missing data.

The current article develops and studies an alternative, analytic (rather than MI based) method for item-level missing data: the two-stage maximum likelihood (TSML). Much like MI does, the TSML method separates the treatment of missing data (Stage 1) from the estimation of the model (Stage 2). It uses the information from Stage 1 in the computation of the standard errors and the model test statistic in Stage 2 to produce correct inferences that account for missing data. The TSML approach to missing data has been shown to be promising in modeling contexts that do not involve composites (Savalei & Bentler, 2009; Savalei & Falk, 2014). The distinction between Stages 1 and 2 of the TSML method is analogous to that between the imputation and the analysis stages in MI. We hypothesize that under the multivariate normal model, the TSML methodology will produce a solution largely equivalent to that obtained from item-level MI based on a large number of imputations.

Other, theoretically suboptimal approaches to treating item-level missing data exist. Two such approaches, scale-level FIML (SL-FIML) and available-case maximum likelihood (ACML), are included in the simulation study described

in this article.[2] SL-FIML declares the entire composite as missing whenever any of its items are missing. This approach can be very inefficient: At the extreme, SL-FIML may even end up with no data if all participants have left at least 1 item of each composite incomplete. The analogous scale-level MI procedure has been shown to perform abysmally relative to item-level MI in terms of efficiency (Gottschall, West, & Enders, 2012). Worse yet, SL-FIML may not always produce consistent parameter estimates: If one of the items within the composite is missing as a function of the value of another item within the same composite, setting the entire composite as missing creates nonignorable missingness.

ACML computes scale scores by averaging all available items and performing maximum likelihood (ML) estimation on the resulting data set. Typically, the resulting data set will be complete, but if it still has missing data (i.e., if some participants left all items on a scale or composite incomplete), FIML estimation can be run. Because ACML is equivalent to imputing the person-level item mean for each missing item score, it too may not always be consistent (Mazza, Enders, & Ruehlman, 2015; Schafer & Graham, 2002). The ACML method is probably the default method employed by practitioners when they encounter item-level missing data.

This article is organized as follows. First, we present the technical details of the normal theory TSML method's adaptation for item-level missing data. Next, we summarize the results of a simulation study comparing the TSML method, SL-FIML, ACML, and item-level MI in the context of an SEM with parcels. The simulation study varied sample size, percentage missing data, type of missing data mechanism, and strength of interitem correlations. We end with limitations and future directions.

Normal Theory Two-Stage Method for Item-Level Missing Data

The two-stage estimator for incomplete data has been in use for a long time but as an ad hoc method. In this form, it involved estimating the population means and covariance matrix from incomplete data under the saturated model (e.g., via the expectation-maximization (EM) algorithm, Dempster, Laird, & Rubin, 1977) and then using these estimates in the complete data ML fit function to fit an SEM (Allison, 2003; Enders & Peugh, 2004; Graham, 2003). This approach has intuitive appeal because it is easy to understand; however, the standard errors and the test statistic are incorrect, as they do not incorporate the uncertainty associated with missing data. Additional corrections are required to make the method statistically valid. Correct standard errors and test statistics to accompany the TSML estimator have been developed for both normal and nonnormal data (Cai, 2008; Cai & Lee, 2009; Savalei & Bentler, 2009; Yuan & Bentler, 2000). These corrections follow the same rationale as the robust corrections developed by Satorra and Bentler (1994), which is essentially an adjustment for loss of efficiency (see Savalei, 2014). When normality is assumed, the

resulting standard errors are largely equivalent to those obtained via special pooling formulas in MI (Asparouhov & Muthén, 2010a, 2010b; Rubin, 1987), if the number of imputations is sufficiently high. When data are nonnormal, TSML with appropriate corrections actually outperforms robust FIML (Savalei & Falk, 2014). The proposed modification of the TSML method to allow item-level missing data is developed in this article under the assumption of multivariate normality; extensions to nonnormal data will be considered in future work.

In the development below, a scale is any type of unit-weighted[3] linear composite, whether a scale score or a parcel. We illustrate the development for two scales but also give the corresponding developments for any number of scales. Let $X = (X_1, X_2, \ldots, X_{p1})'$ represent the random variables on Scale 1, and let $Y = (Y_1, Y_2, \ldots, Y_{p2})'$ represent the random variables on Scale 2. The pooled $p \times 1$ vector of variables from both scales is $Z = (X', Y')'$, where $p = p_1 + p_2$. More generally, the $p \times 1$ vector Z contains items from k scales, where $p = \sum_{i=1}^{k} p_i$. Let $X_c = \sum X_i$ and $Y_c = \sum Y_i$ represent the corresponding scale scores. Let $Z_c = (X_c, Y_c)'$. More generally, Z_c is $k \times 1$.

Stage 1

In this stage, the saturated model is fit to the original items Z using FIML. The obtained saturated estimates of means and the covariance matrix, $\hat{\mu}$ and $\hat{\Sigma}$, are sometimes called the "EM means" and the "EM" covariance matrix (e.g., Enders & Peugh, 2004), even though this terminology confuses the type of estimator with how it was obtained. The $(p^* + p) \times 1$ vector $\hat{\beta}$ contains elements of $\hat{\mu}$ and nonredundant elements of $\hat{\Sigma}$, that is, $\hat{\beta}' = ((\text{vech}\hat{\Sigma})', \hat{\mu}')$ (Magnus & Neudecker, 1999), where $p^* = .5p(p+1)$. We denote the associated observed information matrix by $\hat{A}_\beta$, following the notation of Yuan and Bentler (2000), who also gave an exact asymptotic expression for this matrix when the data are missing completely at random (MCAR; Little & Rubin, 2002). An explicit asymptotic expression for $\hat{A}_\beta$ when data are missing at random (MAR) was given by Savalei (2010).[4] The estimated asymptotic covariance matrix of $\hat{\beta}$ is given by $\hat{\Omega}_\beta = \hat{A}_\beta^{-1}$.

Stage 1a

This additional stage does not occur in the typical TSML procedure (e.g., Savalei & Bentler, 2009) but is needed for the extension of the method to item-level missing data. In this stage, the quantities for raw items from Stage 1 are converted to the corresponding quantities for the scales. To this end, we define a $k \times p$ transformation matrix C such that $Z_c = CZ$. To illustrate, for 2 composites

and 3 items per composite, this matrix is $C = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$. The corresponding saturated estimates of the vector of means and the covariance matrix of the composite variables Z_c are given by $\hat{\mu}_c = C\hat{\mu}$ and $\hat{\Sigma}_c = C\hat{\Sigma}C'$. The $(k^* + k) \times 1$ vector of the saturated model parameters is $\hat{\delta}' = ((\text{vech}\hat{\Sigma}_c)', \hat{\mu}'_c)$, where $k^* = k(k+1)/2$. It is convenient to relate the saturated estimates from Stage 1 to the saturated estimates in Stage 1a: $\hat{\delta} = \begin{pmatrix} D_k^+(C \otimes C)D_p & 0 \\ 0 & C \end{pmatrix}\hat{\beta} = C_{big}\hat{\beta}$, where C_{big} is $(k^* + k) \times (p^* + p)$, D_p is the duplication matrix of order p, and D_k^+ is the Moore–Penrose inverse of the duplication matrix of order k (Magnus & Neudecker, 1999). It follows that the asymptotic covariance matrix of $\hat{\delta}$ is related to the asymptotic covariance matrix of $\hat{\beta}$ as follows: $\hat{\Omega}_\delta = C_{\text{big}}\hat{\Omega}_\beta C'_{\text{big}}$.

Stage 2

In this stage, the researcher's model is fit to the saturated estimates of means and covariance of the composite variables from Stage 1a, Z_c. Let the model representation be $\mu_c = \mu_c(\theta)$, $\Sigma_c = \Sigma_c(\theta)$, where μ_c is the $k \times 1$ vector of population means for the composite variables, Σ_c is the $k \times k$ population covariance matrix of the composite variables, and θ is the $q \times 1$ vector of parameters. The model is fit to data by minimizing the complete data ML fit function:[5]

$$F_{\text{ML}}(\theta) = \text{tr}\{\hat{\Sigma}_c\Sigma_c^{-1}(\theta)\} - \log|\hat{\Sigma}_c\Sigma_c^{-1}(\theta)| + (\hat{\mu}_c - \mu_c(\theta))'\Sigma_c^{-1}(\theta)(\hat{\mu}_c - \mu_c(\theta)) - k,$$

where $\hat{\mu}_c$ and $\hat{\Sigma}_c$ replace what would normally be means and covariance matrix obtained from complete data. The TSML estimates are $\tilde{\theta}$, and the corresponding model-reproduced estimates of means and covariances are $\tilde{\mu}_c = \mu_c(\tilde{\theta})$ and $\tilde{\Sigma}_c = \Sigma_c(\tilde{\theta})$, which are placed in a single vector $\tilde{\delta} = ((\text{vech}\tilde{\Sigma}_c)', \tilde{\mu}'_c)'$.

If the software does not implement the TSML approach, the default standard errors and test statistic from the model run in Stage 2 would be incorrect because the software does not "know" that the fed-in estimates of means and covariances were not based on complete data. The correct estimate of the asymptotic covariance matrix of $\tilde{\theta}$ is given by the "sandwich" estimator:

$$\tilde{\Omega}_{\tilde{\theta}} = (\tilde{\Delta}'\tilde{H}\tilde{\Delta})^{-1}\tilde{\Delta}'\tilde{H}\hat{\Omega}_\delta\tilde{H}\tilde{\Delta}(\tilde{\Delta}'\tilde{H}\tilde{\Delta})^{-1}, \quad (1)$$

where $\tilde{\Delta} = \frac{\partial\delta(\theta)}{\partial\theta'}|_{\theta=\tilde{\theta}}$ is the matrix of model derivatives evaluated at the TSML estimates and $\tilde{H} = \begin{pmatrix} .5D'_k\left(\tilde{\Sigma}_c^{-1} \otimes \tilde{\Sigma}_c^{-1}\right)D_k & 0 \\ 0 & \tilde{\Sigma}_c^{-1} \end{pmatrix}$ is the normal theory weight matrix (the naive "information" matrix) from Stage 2. The quantity $(\tilde{\Delta}'\tilde{H}\tilde{\Delta})^{-1}$ is

the "naive" covariance matrix of parameter estimates that would be produced by default by any software running complete-data ML estimation (Yuan & Bentler, 2000). Equation 1 relies on the estimate of the asymptotic covariance matrix $\hat{\Omega}_{\delta}$ from Stage 1a to produce the right estimates of variability given missing data. Correct standard errors for the TSML estimator $\tilde{\theta}$ are obtained from the diagonal of $\tilde{\Omega}_{\tilde{\theta}}$. For completeness (and because such a discussion can be enlightening), the Appendix discusses what happens when the TSML estimator and Equation 1 are computed on complete data.

To evaluate model fit under normality, the best two-stage test statistic is the residual-based statistic of Browne (1984). Savalei and Bentler (2009) found this statistic to perform very well under normality, and it is the only statistic available for this situation that has an asymptotic χ^2 distribution. Defining model residuals as $\tilde{e} = \hat{\delta} - \tilde{\delta}$, or the difference between the saturated estimates from Stage 1 and the model-implied estimates from Stage 2, the residual-based statistic is given by:

$$T_{\mathrm{RES}} = (N-1)\tilde{e}'(\hat{\Omega}_{\delta}^{-1} - \hat{\Omega}_{\delta}^{-1}\tilde{\Delta}(\tilde{\Delta}'\hat{\Omega}_{\delta}^{-1}\tilde{\Delta})^{-1}\tilde{\Delta}'\hat{\Omega}_{\delta}^{-1})\tilde{e}, \quad (2)$$

where N is sample size. This statistic is referred to a χ^2 distribution with $df = (k^* + k) - q$ degrees of freedom.

Simulation Study: Method

We now summarize the design of a simulation study conducted to evaluate the performance of the new TSML method for item-level missing data relative to three other methods for treating item-level missing data: SL-FIML, ACML, and item-level MI. We hypothesized that TSML and MI would perform similarly, while SL-FIML and ACML would exhibit loss of efficiency and potentially bias in some conditions.

Data Generation

Data on 27 variables were generated from a hierarchical factor model with 9 first-order and 3 second-order factors (see Figure 1). This model was adapted from Coffman and MacCallum (2005), who used it to study the impact of parceling. Models 1 and 2 had identical structure but differed in the strength of first-order factor loadings: They were .3, .4, and .5 (.4 on average) in Model 1 and .6, .7, and .8 (.7 on average) in Model 2. The residual variances were adjusted accordingly, so that each observed variable would have variance 1.

Data were generated in *R* from a multivariate normal distribution. Sample sizes of $N = 200$, 400, or 600 were drawn. One thousand data sets were created in each condition. Once complete data sets were created, nine incomplete data sets were created from each complete data set (for each of the three missingness mechanisms by three percentages of missing data). Fifteen of the 27 variables were to set to

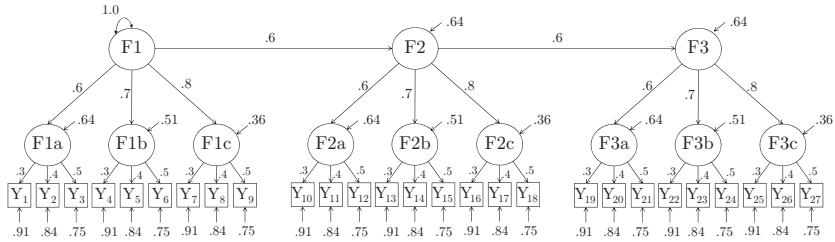

FIGURE 1. *Population Model 1 (used to generate complete data). Model 2 was equivalent to this except the first-order factor loadings were {.6, .7, .8} instead of {.3, .4, .5}, and the residual variances were adjusted so that the total variance of each observed variable was still 1.*

have missing data, and the percentage of missing data per variable was set to be 5%, 15%, or 30%. The variables with missing data were divided into six sets, and the variables within each set were always missing jointly, while missing data across sets were created independently. Thus, the maximum number of missing patterns was $2^6 = 64$. The six sets of incomplete variables were {Y1, Y5, Y9}, {Y10, Y11}, {Y14, Y15, Y16, Y18}, {Y20, Y21}, {Y22, Y24}, and {Y25, Y26}.

Three missing data mechanisms were created: MCAR, MAR linear, and MAR nonlinear. In the MCAR conditions, for a randomly picked row, missing data were created on one of the six sets of variables. This procedure was repeated (with replacement) until the desired percentage of missing data per variable was reached. In the MAR linear conditions, six complete variables (Y2, Y12, Y13, Y19, Y23, and Y27) were used to condition the missing values in each of the six sets of variables. For a randomly picked row, the corresponding set was deleted if the corresponding conditioning variable was greater than 0. This process was repeated (with replacement) until the desired percentage of missingness per variable was reached. In the MAR nonlinear condition, the same procedure was used except missing data were created if the corresponding conditioning variable was greater than .67 in absolute value.

Data Analysis

TSML implementation. To implement Stage 1, the saturated model was run on the full item data using *lavaan* 0.5-18 (Rosseel, 2012) with FIML estimation. The parameter vector $\hat{\beta}$ and its associated asymptotic covariance matrix $\hat{\Omega}_\beta$ (obtained using the vcov() function in *lavaan*) were saved. Stage 1a was implemented using our own R code[6] to obtain $\hat{\mu}_c$ and $\hat{\Sigma}_c$, $\hat{\delta}$, and the asymptotic covariance matrix of $\hat{\delta}$. To implement Stage 2, the analysis model (see Figure 2) was fit to $\hat{\mu}_c$ and $\hat{\Sigma}_c$ using complete data ML estimation in *lavaan*. The sandwich-type standard errors (Equation 1) and the normal theory residual-based statistic (Equation 2)

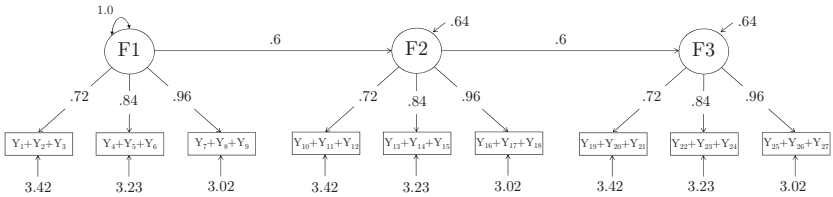

FIGURE 2. *Analysis model (shown with true parameter values for Model 1). Standardized factor loadings for Model 1 are {.363, .423, .484} for each factor. The corresponding true parameter values for Model 2 are as follows: factor loadings are {1.26, 1.47, 1.68} for each factor, residual variances are {4.33, 3.76, 3.10}, and factor variances and regression coefficients are the same as in Model 1. Standardized factor loading values for Model 2 are {.52, .60, .69}. These values were derived algebraically from the corresponding values for the components; the derivations were verified empirically by fitting the analysis model to the population covariance matrices of the composites. The analysis model was fit with (residual) factor variances fixed to their true values, and all loadings, latent regression coefficients, and indicator residual variances freely estimated.*

were computed using our own R code, but we relied on *lavaan*'s internal computeDelta() function to obtain the matrix of model derivatives $\tilde{\Delta}$ necessary for these computations.

SL-FIML implementation. To implement SL-FIML, the original p variables were summed into k composite variables. If any value within a composite was missing, the composite was also set to be missing. The result was an $N \times k$ data matrix with missing values. The analysis model was fit to this data set using FIML estimation in *lavaan*. Parameter estimates, standard errors, and the χ^2 test statistic were extracted from the *lavaan* output.

ACML implementation. To implement ACML, the original p variables were again summed into k composite variables, but individual missing cases were first replaced by the row mean of the remaining complete variables making up the composite. This method is equivalent to taking the mean of the available data within each composite and resulted in an $N \times k$ data matrix with no missing values. The analysis model was fit to this data set using the usual complete data ML estimation in *lavaan*. Parameter estimates, standard errors, and the χ^2 test statistic were extracted from the *lavaan* output.

MI implementation. The MI procedure involved three steps (imputation, analysis, and pooling). In the imputation step, $m = 20$ complete data sets were imputed using the "norm" method in the R package *mice* (van Buuren & Groothuis-Oudshoorn, 2011). The imputation process begins by replacing each missing value in the data set with a randomly drawn observed value from the same

variable to obtain an initial complete data set. Then, a univariate linear regression model is applied to each variable z_i with missing data, such that its missing values are predicted by the remaining variables in the data set. The univariate imputation model is $\dot{z}_i = \dot{\beta}_0 + Z_{\tilde{i}}\dot{\beta}_1 + \dot{\varepsilon}$, where $\dot{z}_i$ is a vector of imputed scores for those individuals with missing data on z_i; $Z_{\tilde{i}}$ is the matrix of scores on $p - 1$ predictor variables (i.e., all other variables in the data set) for those rows on which z_i is missing; $\dot{\varepsilon} \sim N(0, \dot{\sigma}^2)$; and $\dot{\beta}_0$, $\dot{\beta}_1$, and $\dot{\sigma}^2$ are random draws from the full conditional distribution of parameters, given the data (see van Buuren, 2012, for more detail). This process is repeated for each variable with missing data, until every missing value has been imputed. The regression imputation procedure is iterated 20 times,[7] with each iteration using the previous iteration's imputed values in the predictor matrix, $Z_{\tilde{i}}$. The values from the final iteration are saved as a single set of imputed values. The entire procedure is repeated $m = 20$ times to produce 20 imputed data sets.

Each imputed data set was transformed into an $N \times k$ matrix of composite scores. The runMI() function in the *semTools R* package (Pornprasertmanit, Miller, Schoemann, & Rosseel, 2014) was used to fit the analysis model to each imputed data set; this function repeatedly calls *lavaan* and combines the results across imputations. Provided that model estimation in Stage 2 converged to a solution without error messages for at least 3 of the 20 imputed data sets, results were pooled over the converged solutions; otherwise, the replication was categorized as a convergence failure and no further analyses were done. Parameter estimates were averaged across all converged solutions, and standard errors were combined using Rubin's rules (Rubin, 1987). Our choice to average essentially all available imputations is consistent with the defaults in other popular software (e.g., Mplus; Muthén & Muthén, 1998–2015), but we reasoned that three imputations are the bare minimum for obtaining valid inferences (Schafer, 1997; Schafer & Olsen, 1998).

To evaluate model fit, the Meng–Rubin pooled χ^2 was computed using the *semTools* package in *R*; we chose this type of pooled χ^2 because it is also implemented in Mplus, a popular SEM package (Asparouhov & Muthén, 2010a; Meng & Rubin, 1992). In this procedure, a log likelihood is computed for each imputed data set for two fixed-parameter models: an analysis model in which all parameter estimates are fixed to the average parameter values across all imputations and a saturated model in which all parameter estimates are fixed to the average saturated model estimates across all imputations. A likelihood ratio test statistic is computed for each imputation, and these results are averaged over the imputed data sets. The χ^2 pooling procedure converged in all but one of the "eligible" replications (i.e., those in which at least 3 of the 20 imputations converged).

Outcome variables. For each analysis method in each condition, the following measures were collected: convergence and condition code rates, number of

replications containing outliers, parameter bias, efficiency, confidence interval coverage, and rejection rates of the associated test statistic.

An outlier was defined as any parameter estimate of type λ or β (i.e., a factor loading or a latent regression coefficient; see Figure 2) that exceeded 10 in absolute value. Because the inclusion of replications with such large outliers would significantly skew the results, we opted to exclude replications with outliers from the remaining outcome measures. Thus, the results for each method are based on only those replications that converged and exhibited no outliers.

Raw unstandardized bias is the average deviation of the estimate from the true value across replications. Bias was computed separately for each model parameter. For the purposes of presentation, raw bias values were averaged across all parameters of the same type (nine factor loadings, two regression coefficients).[8] While the true values of factor loadings were not all equal in the analysis model (see Figure 2), this summary proved sufficient to draw conclusions about the relative performance of the studied methods. We omitted residual variances and means from the comparison, as they are typically not of direct interest in a factor model.

Efficiency was defined as the empirical standard deviation of parameter estimates in each cell of the design. Because the goal of the study was to compare methods rather than to evaluate the impact of design variables on the efficiency of any given method, we computed efficiency ratios of SL-FIML, ACML, and MI relative to TSML by taking ratios of the corresponding empirical standard deviations. For the purposes of presentation, relative efficiency ratios were averaged across all parameters of the same type.

Coverage of 95% confidence intervals was computed as the number of times out of 1,000 replications (or out of all converged replications with no outliers) that a 95% confidence interval constructed around the parameter estimate using the estimated standard error contained the true value of the parameter. Results were again averaged across type of parameter. Type I error rates were computed as the number of times out of 1,000 (or out of all converged replications with no outliers) that a test statistic produced a p value less than $\alpha = .05$. Because excluding replications that failed to converge can bias Type I error rates (e.g., Yuan & Marshall, 2004), we also computed Type I error rates including such replications.

Simulation Study: Results

Convergence Failures

The number of converged replications with no condition codes differed significantly by model. Under Model 2 (higher factor loadings), convergence was at least 99% in all cells. Under Model 1, convergence was high for the two larger sample sizes ($N = 400$ and 600), averaging 99.3% and never lower than 96.1%. At $N = 200$, average convergence rates were 90.2%, 92.2%, 91.4%, and 97.1% for SL-FIML, ACML, TSML, and MI, respectively. Convergence rates for all

methods but MI were lower with higher proportions of missing data and for more sinister missing data mechanisms, reaching as low as 80.7% (corresponding to SL-FIML). While convergence was highest for MI, this may be due to our liberal definition of convergence (requiring only 3 of the 20 imputations to have converged).

Outliers

As with nonconvergence, outliers were largely limited to Model 1, which had lower population factor loadings. Under Model 2, the vast majority of cells had no outliers, and one cell had a single outlier. Under Model 1, replications with outliers were present in many cells of the design. Most outliers corresponded to latent regression coefficients. Outliers tended to occur more frequently for the larger proportions of missing data and the more difficult missing data mechanisms. The two ad hoc methods, SL-FIML and ACML, had the largest numbers of outliers, while TSML and MI had the lowest. At $N = 200$, the average number of outliers across all conditions of Model 1 was 36, 38, 3, and 2 for SL-FIML, ACML, TSML, and MI, respectively. These averages were under 5 even for the worst performing methods by $N = 400$, however, and essentially 0 by $N = 600$.

Table 1 gives the total number of replications available for the analyses—converged replications that did not produce any outliers—at the two smallest sample sizes. This table makes clear that the interpretation of results for Model 1 at $N = 200$ will be made difficult in some conditions because some methods are missing a considerable number of replications. In fact, based on Table 1 alone, TSML is preferred over SL-FIML and ACML, as it is the analytic method that produces an interpretable solution most often. Table 1 also gives the average number of converged imputations for MI in the replications used in the analysis (i.e., excluding those that have produced fewer than three replications). The number of converged imputations in MI is most affected by sample size and by the type of missing data mechanism but not by percentage of missing data.

Bias in Parameter Estimates

Raw bias was computed separately for each parameter and then averaged across the two types of parameters: nine factor loadings and two latent regression coefficients. Although items in the data generating model had unit variance, parcels created from these items do not have unit variance (see Figure 2). For this reason, raw bias cannot be interpreted on a standardized scale, though it can be used to compare methods.

Factor loadings. There was no bias observed for the MCAR mechanism for any method: Average bias for factor loadings never exceeded .02 in absolute value in any of the MCAR conditions. There was no bias observed for the MAR linear

TABLE 1.
Number of Replications Available for the Analysis for Each Method and Average Number of Converged Imputations for MI

			Model 1					Model 2				
N	%	Mech	SL-FIML	ACML	TSML	MI	Ave	SL-FIML	ACML	TSML	MI	Ave
200	5	MCAR	940	938	945	972	19.0	1,000	1,000	1,000	1,000	20.0
		MAR.lin	943	940	949	982	17.8	1,000	1,000	1,000	1,000	20.0
		MAR.nl	942	934	947	971	**14.4**	1,000	1,000	1,000	1,000	20.0
	15	MCAR	**895**	900	927	977	19.0	1,000	1,000	1,000	1,000	20.0
		MAR.lin	903	907	929	966	17.7	1,000	1,000	1,000	1,000	20.0
		MAR.nl	**872**	**884**	924	977	**14.5**	999	1,000	1,000	1,000	20.0
	30	MCAR	**796**	**853**	**869**	962	19.1	1,000	1,000	1,000	1,000	20.0
		MAR.lin	**781**	**857**	**853**	950	17.5	1,000	1,000	1,000	1,000	20.0
		MAR.nl	**722**	**833**	**854**	962	**14.0**	996	1,000	999	1,000	19.9
400	5	MCAR	997	997	998	1000	19.9	1,000	1,000	1,000	1,000	20.0
		MAR.lin	998	998	997	999	19.8	1,000	1,000	1,000	1,000	20.0
		MAR.nl	997	997	999	1000	19.2	1,000	1,000	1,000	1,000	20.0
	15	MCAR	992	992	994	999	19.9	1,000	1,000	1,000	1,000	20.0
		MAR.lin	994	992	995	999	19.8	1,000	1,000	1,000	1,000	20.0
		MAR.nl	989	990	997	999	19.2	1,000	1,000	1,000	1,000	20.0
	30	MCAR	975	973	985	998	20.0	1,000	1,000	1,000	1,000	20.0
		MAR.lin	953	980	984	999	19.8	1,000	1,000	1,000	1,000	20.0
		MAR.nl	952	981	991	1,000	19.2	1,000	1,000	1,000	1,000	20.0

Note. Data for $N = 600$ are omitted; very few problems occurred at this sample size. Number of replications fewer than 900 and number of average converged imputations fewer than 17 are in bold. N = sample size; "%" = percentage of missing data; "Mech" = missing data mechanism; "Ave" = average number of imputations on which MI results are based; MI = multiple imputation; MAR = missing at random; MCAR = missing completely at random; SL-FIML = scale-level full information maximum likelihood; ACML = available-case maximum likelihood; TSML = two-stage maximum likelihood; nl = nonlinear; lin = linear.

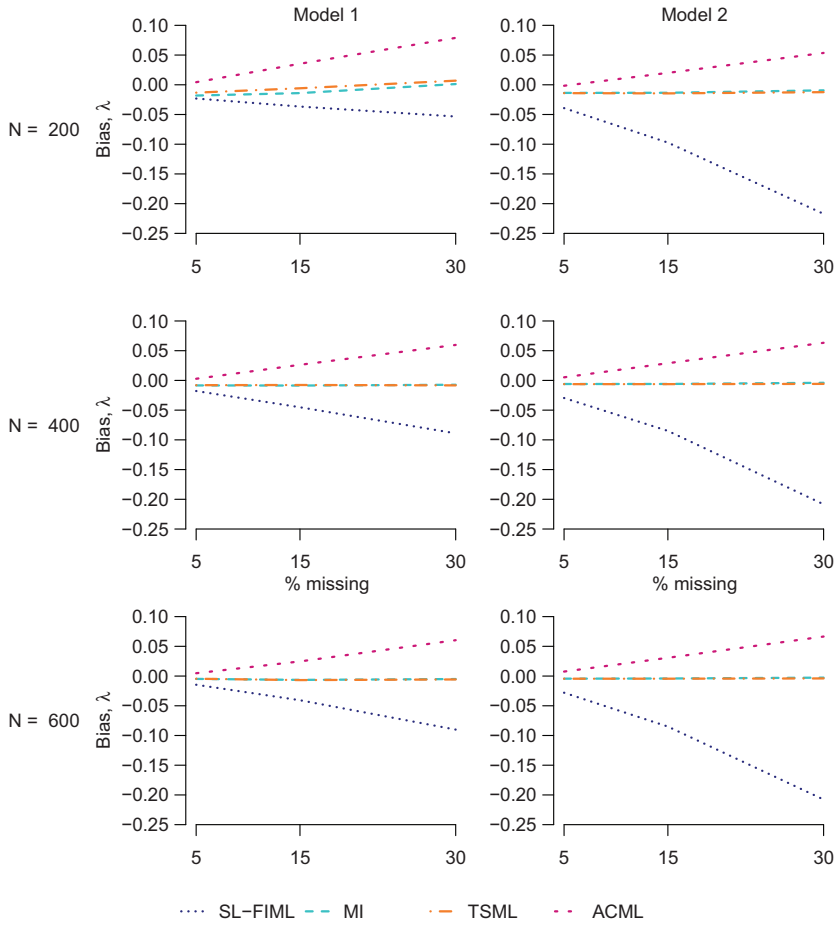

FIGURE 3. *Average bias for factor loading estimates in the missing at random nonlinear conditions.*

mechanism for three of the four methods; average bias for factor loadings never exceeded .02 in absolute for MI, TSML, and ACML. For SL-FIML, small negative bias was observed, particularly in the conditions corresponding to 30% missing data; this bias reached as low as −.08.

Figure 3 plots average bias for factor loadings in the MAR nonlinear conditions. TSML and MI perform best, exhibiting essentially no bias in all conditions. When it comes to the two ad hoc methods, ACML exhibits positive bias, while SL-FIML exhibits negative bias; this bias becomes considerable for greater amounts of missing data and does not decrease with sample size.

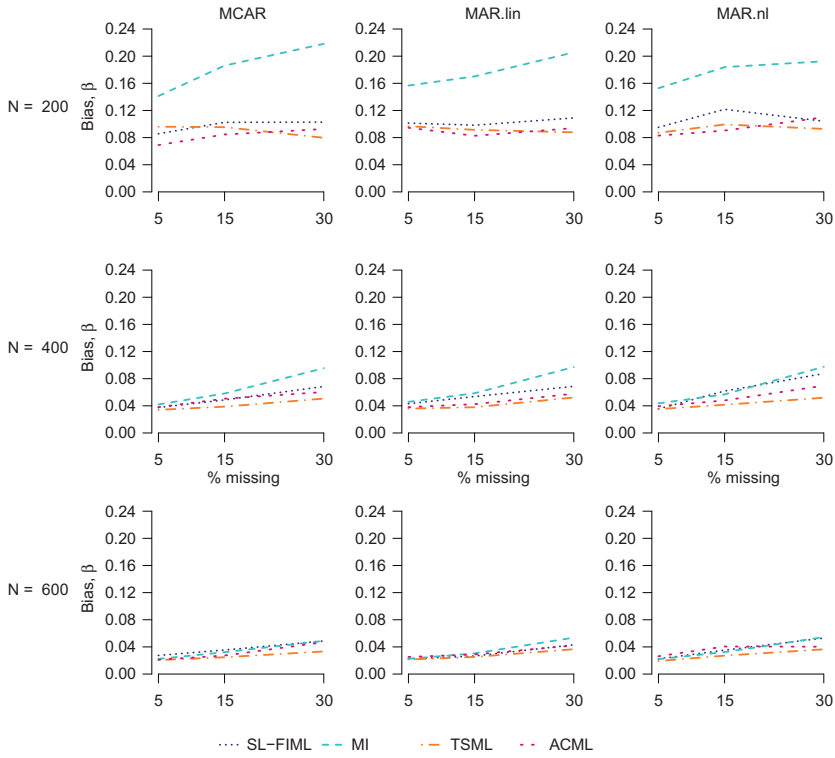

FIGURE 4. *Average bias for latent regression coefficient estimates for Model 1.*

Examining bias for individual factor loadings (not shown) reveals that, within each model, bias is greater for stronger factor loadings. Additionally, conditions where no bias is apparent on average also have little bias for any individual loading (i.e., zero bias is never the result of averaging over positive and negative bias). Investigating bias by latent factor (not shown) reveals that bias is greater for factor loadings of indicators of Factor 3, followed by Factor 2, followed by Factor 1. This finding suggests that the percentage of missing data affects bias (i.e., Factors 2 and 3 had more indicators with missing data than Factor 1) and also that a greater number of missing data patterns may lead to more bias (i.e., one missingness pattern was imposed on indicators of Factor 1, two patterns affected Factor 2, and three patterns affected Factor 3).

Latent regression coefficients. Average bias for latent regression coefficients for Model 1 is plotted in Figure 4. Bias for TSML, ACML, and SL-FIML is similar in most conditions, increasing with proportion of missing data and decreasing with sample size. It is generally small and it does not appear to depend on the

missing mechanism. TSML is almost always the best performing method, while SL-FIML is frequently the worst of the three. Bias for MI is similar to the ML-based methods at the largest sample size but is substantially greater at the two smaller sample sizes, particularly for large amounts of missing data. While this pattern is interesting, it is likely due to the selection effect, as MI has many more replications available for analysis in some of these conditions (see Table 1). To support this explanation, the results for Model 2 (not presented) show negligible bias for all methods in all conditions, never exceeding .025 in absolute value. It is also interesting to note that the bias exhibited by ACML and SL-FIML with factor loadings in the MAR nonlinear conditions (Figure 3) does not appear to propagate to the latent regression estimates.

Efficiency

When the average relative efficiency ratio is larger than 1, the corresponding estimator has, on average (across all parameter estimates of the same type), larger empirical standard deviation than the TSML estimator. When the relative efficiency ratio is less than 1, the corresponding estimator has, on average, smaller empirical standard deviation than the TSML estimator. In the absence of significant bias, efficiency ratios also speak to accuracy of estimation.

Factor loadings. Figures 5 and 6 plot average efficiency ratios for factor loadings for Models 1 and 2, respectively. Because performance is relative to TSML, the line for TSML is horizontal at 1. We first discuss the performance of the two theoretically justified methods. For Model 2, TSML and MI perform very similarly in all study conditions, suggesting that these methods are largely equivalent. For Model 1, TSML and MI exhibit similar performance at the two largest sample sizes. At $N = 200$, MI appears to exhibit an efficiency advantage, producing estimates that are up to 7% less variable than the corresponding TSML estimates. These are also the conditions where considerable selection bias occurs, as MI has quite a few more converged replications (see Table 1). While it is not clear why this results in higher efficiency for MI, this effect is clearly due to selection bias, as it is not present in any other conditions.

The two ad hoc methods, ACML and SL-FIML, both have lower efficiency relative to TSML. The loss of efficiency for SL-FIML follows a fairly constant pattern across sample sizes and missing data mechanisms, depending primarily on the amount of missing data. At the highest rate of missing data, the loss of efficiency for SL-FIML is about 20% under both models. The loss of efficiency for ACML occurs primarily under Model 1, where it is particularly strong under the MAR nonlinear mechanism, while also depending on the proportion of missing data. Interestingly, under Model 2 (with higher factor loadings), ACML performs much more similarly to TSML and MI. Factor loadings (and thus item reliabilities and interitem correlations) are higher under Model 2. With

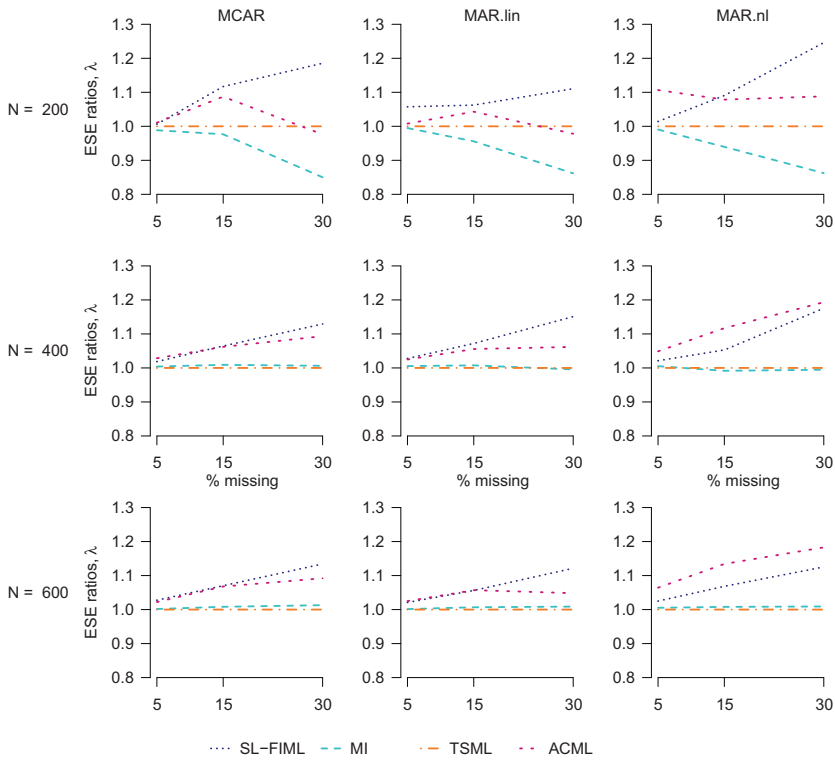

FIGURE 5. *Average efficiency ratios for factor loadings for Model 1.*

increasing factor loadings, the ACML strategy of imputing missing values with the mean of observed values (i.e., averaging over available indicators) becomes more valid, as each observed item is a more reliable stand-in for the whole scale.

Latent regression coefficients. Figures 7 and 8 show the average efficiency ratios for the latent regression coefficients for Models 1 and 2, respectively. For Model 2, the pattern of results is clear: SL-FIML is inferior, particularly at large proportions of missing data, while the other three methods perform similarly. The results for Model 1 are much messier, even at the largest sample size. This is partly due to the selection bias due to the presence of convergence failures and outliers in many conditions. Another reason may be the weaker correlations among the variables in Model 1, so that less information is available in the observed data about the missing data, and the performance of all methods is thus more variable.

However, two patterns can be noted. First, MI no longer performs similarly or better than TSML—in fact, it always performs worse, producing empirical standard errors that can be 20% to 40% larger. Recall that MI estimates were also

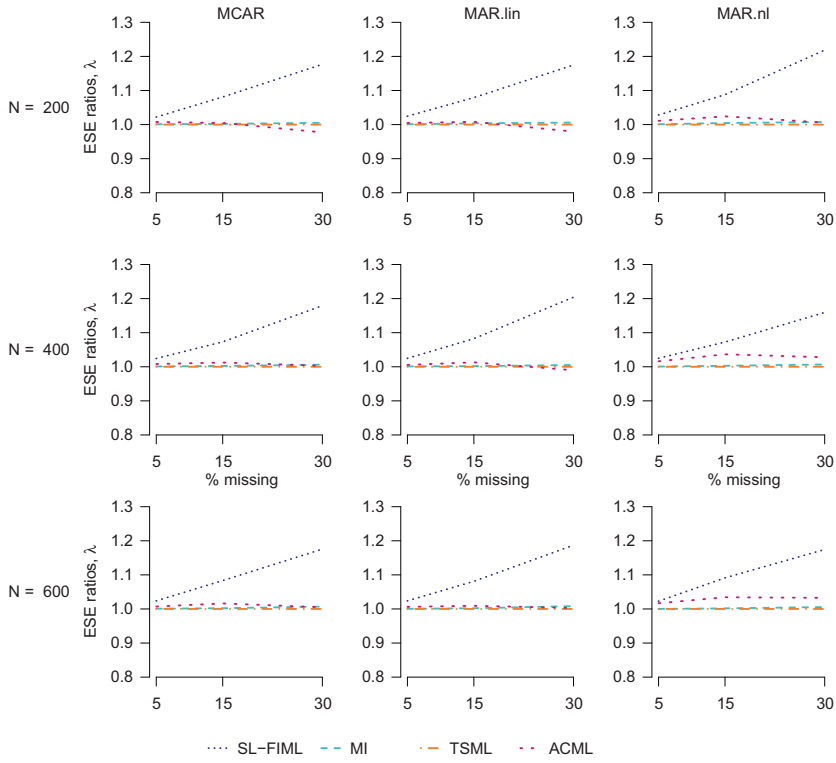

FIGURE 6. *Average efficiency ratios for factor loadings for Model 2.*

biased in these conditions, particularly for large amounts of missing data. Second, SL-FIML and ACML are also generally less efficient than TSML (though at $N = 200$, ACML appears to do better), but the specific patterns across conditions for Model 1 are generally two chaotic to follow.

Coverage

Coverage of 95% confidence intervals is presented in Table 2, averaged across the three missing data mechanisms and parameters of the same type. For factor loadings, coverage is generally good in all conditions for all methods, with the exception of SL-FIML. SL-FIML has low coverage when 30% of data are missing, although this behavior is largely limited to Model 2, and it is worse in the MAR nonlinear conditions. In these conditions, the SL-FIML estimates exhibit large bias (see Figure 3). It is also worth noting that coverage is unreliable to the degree that there are omitted replications in certain conditions: It is difficult or

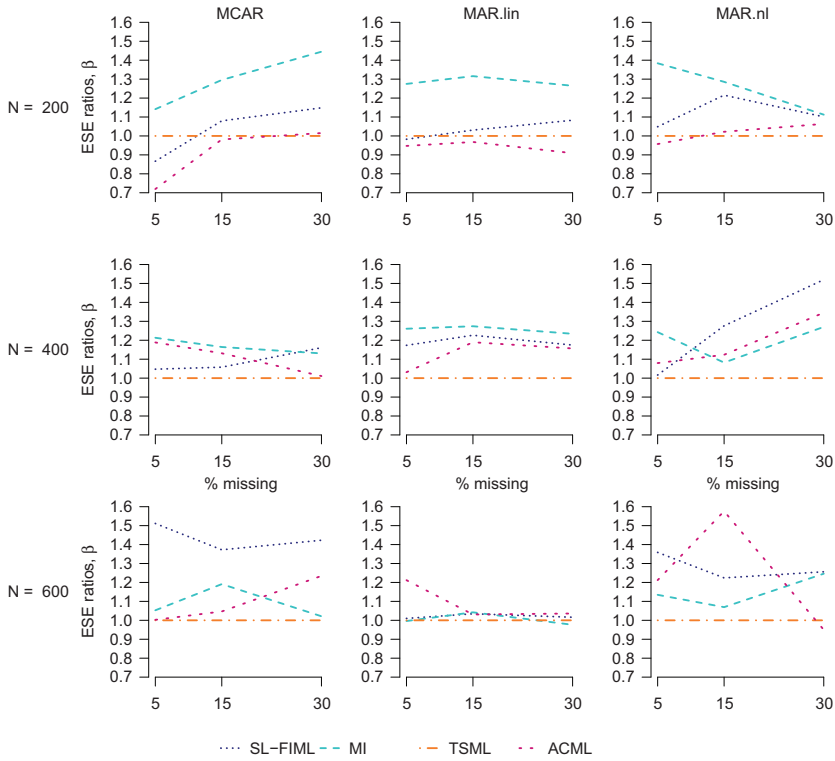

FIGURE 7. *Average efficiency ratios for latent regression coefficients for Model 1.*

impossible to know whether the obtained estimates would have covered the true parameter values or not had those replications converged.

For latent regression coefficients and under Model 2, coverage is excellent in all conditions and for all four methods, including SL-FIML. This pattern of results can be understood by noting that bias of SL-FIML estimates was much lower for regression coefficients than it was for factor loadings. In contrast, high bias in MI regression coefficients was offset by low efficiency, resulting in acceptable coverage. For Model 1, coverage is below 93% for several methods, most notably TSML, in select conditions. These are again the conditions with the most selection bias (see Table 1). However, average coverage never drops below 90%.

Type I Errors

Type I error rates at $\alpha = .05$ are presented in Table 3. Somewhat arbitrarily, we consider the range of 3.5% to 6.5% acceptable. At $N = 600$, all methods

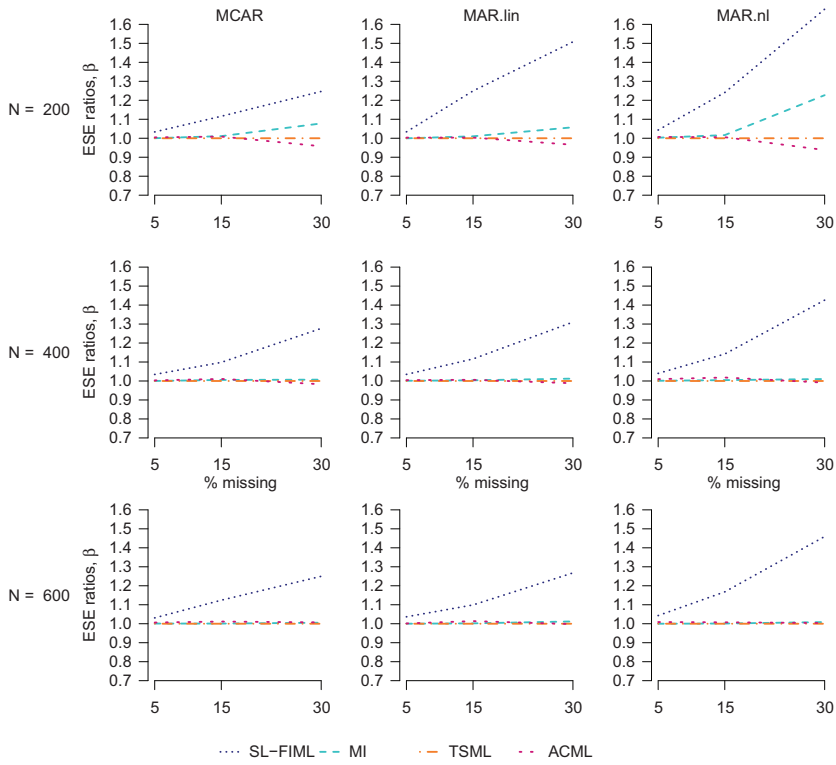

FIGURE 8. *Average efficiency ratios for latent regression coefficients for Model 2.*

displayed acceptable Type I error rates, with only one cell falling outside the acceptable range. At the smallest sample size of $N = 200$, many more problematic rejection rates were observed, particularly with 30% missing data. MI had unacceptably high rejection rates under Model 1, and some rejection rates that were too low under Model 2. SL-FIML and ACML exhibited inflated some rejection rates under Model 2 only. TSML sometimes exhibited rates that were too low under Model 2. Overall, TSML was the best performing method.

Discussion

This article described a new analytic ML-based methodology for handling item-level missing data when the model of interest is based on composites rather than on raw items. The two-stage (TSML) approach, which separates the treatment of missing data from the estimation of the model, has already been established as promising in other missing data contexts (Savalei & Bentler, 2009; Savalei & Falk, 2014). In this article, the TSML method was developed under

TABLE 2.
Coverage of 95% Confidence Intervals, Averaged Across Missing Data Mechanisms and Across Type of Parameter

			Model 1				Model 2			
	N	%	SL-FIML	ACML	TSML	MI	SL-FIML	ACML	TSML	MI
Factor loadings	200	5	95.1	95.1	94.1	94.8	94.1	94.3	94.1	94.1
		15	94.9	95.3	93.8	95.8	93.0	94.3	93.8	94.4
		30	94.1	95.0	93.2	96.6	**89.7**	93.9	93.4	95.0
	400	5	94.7	94.7	94.0	94.3	94.3	94.6	94.4	94.4
		15	94.3	94.6	93.9	94.9	93.1	94.2	94.4	94.6
		30	93.7	94.2	94.0	96.0	**86.9**	93.2	94.0	94.7
	600	5	94.6	94.5	94.4	94.4	94.8	94.9	95.0	95.0
		15	94.2	94.3	94.1	94.7	**92.7**	94.6	95.0	95.1
		30	93.1	93.8	94.0	95.2	**84.8**	**92.7**	94.8	95.2
Latent regression coefficients	200	5	94.0	93.9	**92.7**	93.7	94.5	94.5	94.4	94.5
		15	**92.6**	93.1	**91.5**	94.1	94.3	94.5	94.2	94.6
		30	**92.3**	**92.1**	**90.3**	95.3	93.5	94.3	93.5	94.8
	400	5	94.0	94.0	93.3	93.6	95.0	94.8	94.5	94.5
		15	93.9	93.3	**92.7**	93.7	95.0	95.2	94.6	94.8
		30	**92.3**	**92.2**	**91.8**	94.1	94.7	95.2	94.4	94.7
	600	5	94.6	94.2	94.1	94.2	95.5	95.8	95.5	95.5
		15	94.2	93.6	94.1	94.5	95.5	95.9	95.4	95.7
		30	93.2	93.5	93.1	94.3	94.6	95.8	95.4	95.6

Note. Values below 93% are in bold. Coverage is based on all converged replications with no outliers and is averaged across each type of parameter estimate and across the missing data mechanisms. MI = multiple imputation; SL-FIML = scale-level full information maximum likelihood; ACML = available-case maximum likelihood; TSML = two-stage maximum likelihood.

the assumption of the multivariate normal distribution for the items. Because in Stage 1 the TSML methodology uses FIML estimation under the saturated model, the output of this stage is equivalent to that of item-level MI under the normal model, at least for a large number of imputations. Prior to the current development, item-level MI was the only available method to treat item-level missing data appropriately when fitting a model to composites. The newly developed TSML method is thus a viable analytic alternative to item-level MI and, once implemented in software, should be appealing to researchers who prefer ML-based methods over running MIs.

Prior to conducting the simulation study described in this article, we predicted that item-level MI and the new TSML method would have increasingly similar performance in terms of bias, efficiency, and coverage, as sample size grew large, because they are essentially asymptotically equivalent (for an infinite number of imputations). This prediction was largely confirmed.

TABLE 3.
Type I Error Rates at $\alpha = .05$

			Model 1				Model 2			
N	%	Mech	SL-FIML	ACML	TSML	MI	SL-FIML	ACML	TSML	MI
200	5	MCAR	5.1	4.8	4.2	6.4	6.0	**6.7**	**2.9**	4.8
		MAR.lin	6.5	5.6	5.0	**9.6**	5.8	**6.8**	4.3	3.9
		MAR.nl	5.3	3.5	6.3	**9.5**	**6.9**	**6.7**	5.9	**2.8**
	15	MCAR	4.7	4.6	3.6	6.5	6.5	5.9	**3.4**	4.9
		MAR.lin	4.7	5.5	3.6	**6.8**	6.3	6.2	3.6	4.1
		MAR.nl	5.4	4.0	5.6	**8.2**	**7.7**	**7.2**	6.5	**3.1**
	30	MCAR	5.6	4.8	4.0	**6.6**	**6.7**	6.3	**3.4**	4.8
		MAR.lin	5.4	5.3	4.5	**8.0**	6.0	5.7	4.3	4.0
		MAR.nl	4.3	6.2	5.9	**10.5**	**7.7**	**7.2**	**6.9**	**2.7**
400	5	MCAR	5.8	5.6	4.6	5.3	6.1	6.1	4.6	5.6
		MAR.lin	6.2	5.6	5.2	6.0	4.8	6.5	5.4	5.7
		MAR.nl	5.6	5.3	5.8	**8.5**	**6.9**	5.9	**6.9**	4.3
	15	MCAR	5.3	5.5	5.2	5.6	5.8	6.5	4.8	5.8
		MAR.lin	4.7	6.1	5.4	6.2	6.1	5.2	4.3	4.3
		MAR.nl	5.6	5.1	5.1	6.1	**6.6**	5.4	6.0	3.5
	30	MCAR	5.0	5.3	4.4	5.5	5.6	6.4	4.9	5.6
		MAR.lin	5.9	5.8	4.9	6.0	5.7	**6.6**	5.7	5.5
		MAR.nl	5.0	5.7	5.8	**8.3**	5.5	5.4	6.2	4.5
600	5	MCAR	5.3	5.6	4.6	5.2	5.3	5.7	4.5	5.5
		MAR.lin	5.3	6.2	4.7	5.1	5.0	5.4	4.5	4.5
		MAR.nl	5.6	5.1	5.3	4.4	4.7	5.4	5.2	3.9
	15	MCAR	5.8	5.5	4.8	5.5	5.4	6.0	5.1	5.3
		MAR.lin	5.1	5.2	5.2	5.7	5.6	5.6	4.8	4.2
		MAR.nl	5.2	4.9	5.8	6.2	**7.4**	5.5	5.4	4.4
	30	MCAR	4.9	5.0	4.4	5.0	5.4	5.8	4.7	5.2
		MAR.lin	4.9	5.6	4.5	5.2	5.5	5.6	5.1	5.1
		MAR.nl	5.2	4.2	5.0	5.1	5.3	5.0	5.2	3.6

Note. Values outside of the 3.5–6.5% range are in bold. Type I errors are based on all converged replications with no outliers. Including replications with outliers does not change the results in any appreciable way. MI = multiple imputation; MAR = missing at random; MCAR = missing completely at random; SL-FIML = scale-level full information maximum likelihood; ACML = available-case maximum likelihood; TSML = two-stage maximum likelihood; nl = nonlinear; lin = linear

In addition to TSML and MI, we have included in our simulation study two other, theoretically suboptimal approaches to treating item-level missing data: SL-FIML and ACML. These approaches are "ad hoc" in the sense that they can lack the property of consistency when applied to ignorable item-level missing data. Even when consistent, they can be very inefficient. SL-FIML, which sets

the entire composite to missing if any of its items are missing, exhibited the largest amount of bias in the present study, particularly with MAR nonlinear data, as well as the greatest loss of efficiency for all types of missing data. ACML, in which all available items are averaged to obtain composite scores, exhibited some bias under the MAR nonlinear mechanism and also exhibited some loss of efficiency with this type of data.

While our results for SL-FIML and ACML were consistent with our theoretical predictions, the results for ACML in particular may seem sufficiently robust to warrant its continued use by practitioners when they encounter item-level missing data. In particular, the method had little bias in many of the study's conditions as well as good coverage and Type I error rates. We want to stress that this conclusion would be mistaken. We did not set out to "break" ACML when designing our study conditions (in fact, the original study design did not include ACML in the group of methods) but rather to test the empirical performance of the newly developed TSML approach and to compare it with MI. The good performance of ACML is likely an artifact of the study design. Items in our generating model had equal variance and equal means. If items on which missingness was imposed had different variance or different means than fully observed items, ACML would result in composites with correspondingly different variance or means than the true (unobserved) composite scores and would thus lose consistency. This would happen even in the MCAR case. For instance, suppose the item with the highest mean had 50% MCAR missingness, the ACML composite scores for those 50% of participants would be lower and less variable than if those scores were not missing; similarly, if the item with the greatest variance had 50% MCAR missing, the ACML composite scores would have reduced variance. The ACML and the SL-FIML approaches are not recommended and should be avoided (see also Mazza et al., 2015).

The current study assumed continuous normally distributed data at the item level. It is fairly straightforward to extend the TSML approach to continuous nonnormal data (e.g., Savalei & Falk, 2014; Yuan & Bentler, 2000; Yuan & Lu, 2008). The extension to categorical data in Stage 1 is less straightforward, since analytic methods for treating incomplete categorical data are not yet available. However, research with complete data suggests that categorical variables can safely be treated as continuous once the items have five to seven categories (which is very common in behavioral research) and in some cases as few as four (Rhemtulla, Brosseau-Liard, & Savalei, 2012). Moreover, studies of imputation approaches for categorical missing data find that imputation under the normal model does as well as or better than categorical imputation approaches (Finch, 2010; Wu, Jia, & Enders, 2015)—even for binary and three-category data. The work of Wu, Jia, and Enders (2015) is particularly relevant as these authors study MI of categorical items when the model is at the composite level (and composites are then treated as continuous)—that is, the exact situation the TSML method was developed for. These authors found that, across all study conditions,

imputation under the normal model and a state-of-the-art imputation approach for categorical data called the "latent variable model imputation" approach (Asparouhov & Muthén, 2010a, 2010b) perform very similarly to each other and beat all other imputation approaches considered. Recall that imputation under the normal model has been shown to be largely equivalent to TSML in this article. While these studies have used simple regression models, the model itself is likely not relevant, as the goal of imputation is to produce high-quality Stage 1 estimates, to which any subsequent model could be fit. Nonetheless, to test this conjecture, we plan to compare the TSML approach to categorical imputation approaches in future research.

One of the study's limitations is that we evaluated the performance of the TSML approach on a single type of model, which was a full SEM. In future studies, we plan to evaluate the TSML approach with other types of models, such as path analysis and regression models with scale scores—another common application of composite scores. A limitation of the TSML approach is that it is currently not available in software, though we make our sample R code available at osf.io/yx7bf/. In the future, we plan to create an R package that performs this approach or add this method to an existing package, such as *lavaan*.

Appendix

Complete Data Special Case

It can be instructive to work through happens when complete data are analyzed using an incomplete data routine, as the results can differ from the corresponding complete data analysis. Results that are asymptotically the same may differ in small samples. For instance, when the FIML estimator is applied to complete data, the results may not match due to differences in the sample size multiplier used (some programs use $N - 1$ for complete data and N for incomplete data), type of information matrix used (the default in many programs is to use expected information for complete data and observed information for incomplete data), whether information matrices are obtained using exact asymptotic formulas or from the derivatives of the likelihood, and so on. Below we summarize what happens when the TSML implementation described in this article is applied to complete data.

When data are complete, the TSML estimates are the same as ML estimates. The asymptotic covariance matrix of these estimates is given by $(\Delta' H \Delta)^{-1}$, where Δ and H are the matrix of model derivatives and the normal theory weight matrix, respectively, both evaluated at the true parameter values. For complete data, the estimated asymptotic covariance matrix of the TSML estimates given by Equation 1 reduces to $(\Delta' H \Delta)^{-1}$ asymptotically because $\hat{\Omega}_\delta \to H^{-1}$.

However, in finite samples, differences between running the ML routine and the TSML routine on complete data are possible, due to the fact that the estimate

$\hat{\Omega}_{\delta}$ from Stage 1a is not going to be exactly equal to the computation of $\tilde{H}^{-1}$ from Stage 2. Stage 1 uses saturated estimates, while Stage 2 uses structured estimates for the covariance matrix in the expression for $\tilde{H}^{-1}$. It is possible to evaluate H at the saturated estimates in Stage 2, but we did not evaluate this computational option (with complete data, then, the resulting standard errors will be equivalent to generalized least squares (GLS) standard errors and not ML standard errors). Second, Stage 1 uses observed standard errors, since it assumes ignorable missing data and requires this specification for consistency (Savalei, 2010). Stage 2, on the other hand, runs the complete data ML routine and therefore uses expected standard errors. While it is possible to request observed information in Stage 2, in our informal simulations, we found that expected information works better because observed information creates confidence intervals that are too wide in smaller sample sizes.

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) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by grant RGPIN-2015-05251 from the Natural Sciences and Engineering Research Council to V. Savalei and grant FP7-PEOPLE-2013-CIG-631145 from the European Research Council to M. Rhemtulla.

Notes

1. The full information maximum likelihood solution at the item level is of course possible, and the resulting estimates can be transformed to the corresponding estimates for the composites. However, this solution requires the specification of the correct model for the items, which is typically not what the researcher interested in the composite-level model wishes to do. Additionally, it is possible that the composite-level model specified by the researcher holds, while the proposed full item-level model does not hold, so the results from such an analysis will answer fundamentally a different question.
2. Commenters have suggested that there may additionally be a model-based solution to item-level missing data, that is, a single-stage solution in which composites are modeled as latent factors, with items as either reflective or causal indicators with fixed loadings. The causal indicator approach works as long as composites are exogenous variables in the SEM. With endogenous composites, however, such models are either not identified (causal indicator models) or lead to completely different results than the composite-level model (reflective indicator models).

3. The incorporation of nonunit weights is straightforward in the proposed methodology.
4. Structural equation model programs use the second derivatives of the log likelihood in place of these exact asymptotic expressions, but these computational differences are typically not consequential, although research is still needed in this area.
5. Actually, any fit function can be used in Stage 2 of the two-stage method (e.g., least squares), with straightforward adjustments to the computations that follow. However, using complete data ML is likely to produce a more efficient estimator (and, if there is no missing data, an asymptotically fully efficient estimator).
6. Sample *R* code is available for download at osf.io/yx7bf/.
7. In the runMI function specification, maxit = 20.
8. Full summary data are available for download at osf.io/yx7bf/.

References

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology*, *112*, 545–557.

Arbuckle, J. L. (1996). *Full information estimation in the presence of incomplete data*. Mahwah, NJ: Lawrence Erlbaum.

Asparouhov, T., & Muthén, B. (2010a). *Chi-square statistics with multiple imputation, Version 2* (Mplus Technical Report). Retrieved from http://www.stat.model.com

Asparouhov, T., & Muthén, B. (2010b). *Multiple imputation with Mplus, Version 2* (Mplus Technical Report). Retrieved from http://www.statmodel.com

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance-structures. *British Journal of Mathematical & Statistical Psychology*, *37*, 62–83.

Cai, L. (2008). SEM of another flavour: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical & Statistical Psychology*, *61*, 309–329.

Cai, L., & Lee, T. (2009). Covariance structure model fit testing under missing data: An application of the supplemented EM algorithm. *Multivariate Behavioral Research*, *44*, 281–304.

Coffman, D. L., & MacCallum, R. C. (2005). Using parcels to convert path analysis models into latent variable models. *Multivariate Behavioral Research*, *40*, 235–259.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedure. *Psychological Methods*, *6*, 330–351.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B-Methodological*, *39*, 1–38.

Enders, C. K., & Peugh, J. L. (2004). Using an EM covariance matrix to estimate structural equation models with missing data: Choosing an adjusted sample size to improve the accuracy of inferences. *Structural Equation Modeling*, *11*, 1–19.

Finch, W. H. (2010). Imputation methods for missing categorical questionnaire data: A comparison of approaches. *Journal of Data Science*, *8*, 361–378.

Gottschall, A., West, S. G., & Enders, C. K. (2012). A comparison of item-level and scale-level multiple imputation for questionnaire batteries. *Multivariate Behavioral Research*, *47*, 1–25.

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling*, *10*, 80–100.

Larsen, R. (2011). Missing data imputation versus full information maximum likelihood with second-level dependencies. *Structural Equation Modeling*, *18*, 649–662.

Lawrence, A., & Lee, T. (2015, July). A comparison of maximum likelihood and multiple imputation for structural equation models with missing data. Paper presented at the annual International Meeting of the Psychometric Society, Madison, WI. Abstract retrieved from http://conferencing.uwex.edu/conferences/ps2014/docs/IMPS_Abstracts_Web.pdf

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.

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

Magnus, J. R., & Neudecker, H. (1999). *Matrix differential calculus with applications in statistics and econometrics* (Rev. ed.). Chichester, England: John Wiley.

Mazza, G. L., Enders, C. K., & Ruehlman, L. S. (2015). Addressing item-level missing data: A comparison of proration and full information maximum likelihood estimation. *Multivariate Behavioral Research*, *50*, 504–519.

Meng, X., & Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika*, *79*, 103–111.

Muthén, L. K., & Muthén, B. O. (1998–2015). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

Pornprasertmanit, S., Miller, P., Schoemann, A., & Rosseel, Y. (2014). *semTools: Useful tools for structural equation modeling* (R package version 0.4-6). Retrieved from http://CRAN.R-project.org/package=semTools

Rhemtulla, M., Brosseau-Liard, P., & Savalei, V. (2012). How many categories is enough to treat data as continuous? A comparison of robust continuous and categorical SEM estimation methods under a range of non-ideal situations. *Psychological Methods*, *17*, 354–373.

Rosseel, Y. (2012). lavaan: An R Package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: John Wiley.

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.

Savalei, V. (2010). Expected vs. observed information in SEM with incomplete normal and nonnormal data. *Psychological Methods*, *15*, 352–367.

Savalei, V. (2014). Understanding robust corrections in structural equation modeling. *Structural Equation Modeling*, *21*, 149–160.

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.

Savalei, V., & Falk, C. (2014). Robust two-stage approach outperforms robust FIML with incomplete nonnormal data. *Structural Equation Modeling*, *21*, 280–302.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London, England: Chapman & Hall.

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

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, *33*, 545–571.

van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL: Chapman & Hall/CRC.

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

Wu, W., Jia, F., & Enders, C. (2015). A comparison of imputation strategies for ordinal missing data on Likert scale variables. *Multivariate Behavioral Research*, *50*, 484–503.

Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, *30*, 165–200.

Yuan, K.-H., & Lu, L. (2008). SEM with missing data and unknown population distributions using two-stage ML: Theory and its application. *Multivariate Behavioral Research*, *43*, 621–652.

Yuan, K.-H., & Marshall, L. (2004). A new measure of misfit for covariance structure models. *Behaviormetrika*, *31*, 67–90.

Yuan, K.-H., Yang-Wallentin, F., & Bentler, P. M. (2012). ML versus MI for missing data with violation of distribution conditions. *Sociological Methods & Research*, *41*, 598–629.

Authors

VICTORIA SAVALEI is an associate professor of quantitative psychology at the University of British Columbia, 2136 West Mall, Vancouver, BC, Canada V6T 1Z4; vsavalei@psych.ubc.ca. Her research interests include latent variable modeling, especially structural equation modeling (SEM), development of new statistical methods to handle incomplete data, nonnormal data, and categorical data.

MIJKE RHEMTULLA is an assistant professor of quantitative psychology at the University of California, Davis, 1 Shields Avenue, Davis, CA 95616; mrhemtulla@ucdavis.edu. Her research interests include structural equation modeling, methods and designs for missing data, and alternative measurement models for psychological constructs.

Manuscript received February 20, 2016
First revision received May 30, 2016
Second revision received December 5, 2016
Accepted December 13, 2016