STRUCTURAL EQUATION MODELING: A MULTIDISCIPLINARY JOURNAL
2021, VOL. 28, NO. 4, 506–517
https://doi.org/10.1080/10705511.2020.1850289

Is Item Imputation Always Better? An Investigation of Wave-Missing Data in Growth Models

Juan Diego Vera and Craig K. Enders

University of California Los Angeles

ABSTRACT

Questionnaire data present challenges, as a missing item of a multi-item scale would lead to a total missing scale. A researcher applying multiple imputation to an incomplete multi-item questionnaire can impute the incomplete items prior to computing scale scores or impute the scale score entirely. Methodologist have favored item-level imputation because it greatly enhances precision in comparison to scale-level imputation; however, this benefit in precision might not translate into longitudinal data studies where entire questionnaire batteries are missing. We investigated the performance of item- and scale-level imputation model and found that item-level imputation did not produce a precision advantage in estimating any of the growth model parameters and scale-level showed better precision in estimating the slope variance parameter than item-level imputation.

KEYWORDS

Longitudinal studies; missing data; multiple imputation; attrition; questionnaires

Common recommendations for missing data methods include multiple imputation (MI) and full information maximum likelihood (FIML) estimation. These methods work well when variables are missing at random (MAR), which is a mechanism that occurs when the tendency for missing data on one or more variables is explained by other observed variables (Little & Rubin, 2019; Rubin, 1976). FIML is a single-stage analysis, where the population parameters with the highest likelihood of producing the sample data are directly estimated from the observed data; missing values are not imputed, and the estimator uses a normal distribution to infer the missing parts of the data at each iteration (Arbuckle, 1996). In comparison to multiple imputation, FIML is often more efficient in small samples (Yuan et al., 2012), however both multiple imputation and FIML are asymptotically equivalent when applying the same assumptions with the same set of input variables (Collins et al., 2001).

The focus of this study is on multiple imputation, specifically multiple imputation for multiple-item questionnaires. Multiple imputation has three phases that the researcher has to follow: the imputation phase, analysis phase, and pooling phase. In the imputation phase, new values are drawn from a distribution of probable scores, then in the analysis phase, the substantive model is estimated from each of the imputed datasets. In the pooling phase, the parameter estimates and standard errors across all the imputed datasets are combined to obtain a set of results, accounting for both the between- and within-imputation variations (i.e., missing data uncertainty and sampling error, respectively). More detailed description of multiple imputation and the pooling process is available in the literature (Enders, 2010; Schafer, 1997; Schafer & Graham, 2002; Van Buuren, 2012).

The specific multiple imputation variant that we examined in this study is fully conditional specification (FCS; Van Buuren et al., 2006). The fully conditional specification approach specifies a different imputation model per each variable and imputes variables one at a time. The FCS approach draws replacement values from a series of univariate distributions that condition on a set of regression parameters, residuals, and complete and previously imputed variables (Van Buuren, 2007, 2012; Van Buuren et al., 2006). A Bayesian estimation sequence is used for every incomplete variable to generate parameters from a model with an incomplete variable regressed on all remaining variables, and the imputation step then uses the model parameters to create a distribution from which it draws plausible replacement values. A detailed description of FCS imputation is available in the literature (Van Buuren, 2007, 2012). For the problem we are studying, the choice between FCS and joint model imputation (Schafer, 1997, 2001; Schafer & Olsen, 1998) is arbitrary because the two methods are formally equivalent if the sample size is large enough to negate the effect of the prior distribution (Hughes et al., 2014).

The current study will focus on missing data in questionnaires that are composed of multiple multi-item scales where items are summed to create a scale score. Questionnaire data present challenges with missing data, as a missing item of a multi-item scale would lead to ambiguity about how to compute the composite score. A common method for handling missing item responses in applied research is to compute prorated scale scores by averaging the available items. However, this method is suboptimal because it assumes that all inter-item correlations within a scale are the same, all means are the same, and the mechanism is missing completely at

CONTACT Juan Diego Vera jdverajones@g.ucla.edu Department of Psychology, University of California, Los Angeles, Franz Hall, 502 Portola Plaza, Los Angeles, CA 90095.

Supplemental data for this article can be accessed on the publisher's website.

random (Mazza et al., 2015; Schafer & Graham, 2002). Further, this approach cannot be applied to the longitudinal problems where participants skip entire waves and thus have no item responses to average.

Beyond prorated scale scores, a variety of specialized procedures have been proposed for incomplete questionnaire items (Ginkel et al., 2007; Van Ginkel et al., 2010; Vermunt et al., 2008), but we focus on standard multiple imputation because most researchers would likely use this approach given its widespread availability in software. A researcher applying multiple imputation to an incomplete multi-item questionnaire can impute the incomplete items prior to computing scale scores, then sum the imputed item scores and use the total sum as the scale score. Another option is to treat the scale score as missing for participants that have one or more missing item responses, and then impute the scale score. We refer to these approaches as item-level and scale-level imputation, respectively.

To date, previous research has favored item-level imputation because it greatly enhances precision in comparison to scale-level imputation. For example, a study by Gottschall et al. (2012) found that, in certain conditions, the reduction in power from imputing at the scale-level required a 75% increase in sample size to reach the same sampling variance as item-level imputation. They also found that as the number of items within a scale increases, the benefits of item-level imputation (e.g., better power) increase because imputation leverages the strong correlations with other items from the same scale. Similarly, a study by Mazza et al. (2015) found that addressing missing data at the item level rather than the scale level dramatically improves precision. Although this study did not consider imputation, they used an analogous maximum likelihood procedure that uses Graham's (2003) saturated correlates model to incorporate questionnaire items as auxiliary variables.

In the context of longitudinal data, Eekhout et al. (2015) focused on a similar approach that used items as auxiliary variables in a latent growth model with incomplete scale scores. Similar to the studies previously mentioned, the researchers found that including auxiliary item information into the model when item scores are missing improves results compared to not including item-level information. An important feature of this study is that it assumes each individual provides responses to a subset of the items at each wave, such that the entire scale is never missing. This is an important point of distinction with our study. In contrast to the previous cited studies, Rombach et al. (2018) found that scale-level imputation performs better than item-level at a small sample size conditions (e.g., sample size of 100 and 200), where item-level imputation might not be feasible given issues of non-convergence. However, item-level imputation was generally more accurate than scale-level for high proportions of item-nonresponse, in the conditions where convergence was not an issue.

Importantly, studies mentioned previously do not investigate the performance of scale and item-level imputation when all items in a scale are missing. This pattern of missingness is common, if not typical, in longitudinal data, where all item responses may be missing because a participant skips a data collection wave. Not only is the number of missing observations for a given participant substantially larger when this pattern occurs, imputation can no longer borrow strength from items assessed at the same wave. Instead, the procedure must impute items (or scales) from items (or scales) collected at previous or later waves. As a result, there will be a relatively large number of missing values within a given participant and a potentially substantial decrease of observed information to support the item-level imputation model. Despite the fairly consistent advice to impute at the item level prior to computing scale scores, we anticipate that scale-level imputation could produce similar or more accurate predictions in this context because it effectively leverages the same information as item-level imputation (i.e., inter-wave versus intra-wave correlations), albeit with far fewer parameters. The purpose of the present study is to determine the best way to handle missing item responses where entire waves are missing. To do so, we examine the study features (e.g., sample size, questionnaire length, etc.) that impact the effectiveness of item-level or scale-level imputation in longitudinal studies where an entire questionnaire battery is missing at one or more measurement occasions. Guided by previous research, we carefully selected conditions expected to be influential on the relative performance of item-level or scale-level imputation based on previously published simulation studies, and we used computer simulation to examine this issue in the context of a latent growth curve analysis.

Methods

A series of simulation studies were designed to explore the differences between scale-level and item-level imputation. Specifically, we implemented a full factorial design that was comprised of one within-subject factor (the two imputation methods) and four between-subject factors: (a) number of items per scale, (b) magnitude of inter-item correlations, (c) rate of missing item data across waves, and (d) sample size. This resulted in a total of 36 simulated conditions and 2000 data sets within each between-subjects design cell. We selected the between-subject factors because previous research leads us to anticipate that they would be influential on the relative performance of item-level or scale-level imputation (Eekhout et al., 2015; Gottschall et al., 2012; Mazza et al., 2015). We simulated longitudinal data with five measurement waves, with one scale score per wave. The simulation used either five, 10, or 15 items per scale to mimic the range of items in a typical questionnaire design, and the inter-item correlations were varied between .30 and .50. The simulation induced two missing data rates to mimic scenarios with high and low attrition (explained below). The last manipulated subject factor was sample size, which was set at 250, 500, or 1000 artificial scores.

Data generation model description

The data were generated with a higher-order latent growth curve analysis model where scales are represented as repeated latent factors and items are represented as lower-level indicator variables (Bollen & Curran, 2005; Grimm et al., 2016; Little, 2013; Newsom, 2015). The data were ultimately analyzed by summing the item responses and

submitting the scale scores to a conventional latent growth model. The starting point of constructing a latent curve model (LCM) within structural equation modeling (SEM) framework is very similar to a two-factor multiple-indicator confirmatory factor analysis (CFA) model (Jöreskog, 1969; Mcardle & Epstein, 1987; Meredith & Tisak, 1990). A latent factor η is specified for each set of indicators at time point t, and the repeated measurements of η function as indicators of the higher order growth factors. These higher order latent factors represent the linear trajectory model, which includes a latent intercept and the slope for each individual (Bollen & Curran, 2005; Grimm et al., 2016; Little, 2013; Newsom, 2015).

The latent curve measurement model has multiple indicators for each repeatedly measured latent variable η, as follows

$$y_{itq} = \nu_{tq} + \gamma_{tq}\eta_{it} + \varepsilon_{itq} \quad (4)$$

where y_{itq} is score for the q^{th} item at time t for case i, ν_{tq} is the intercept of q^{th} item, and t^{th} time period or wave. Further, γ_{tq} is the factor loading of the q^{th} item in the t^{th} wave, η_{it} is the latent variable for case i at time t, and ε_{itq} is the error of the q^{th} item in the t^{th} wave, for case i.

The higher-order model uses the repeatedly measured latent factors as indicators of the higher-order intercept and slope latent factors. The model for the repeated latent variables is

$$\eta_{it} = \alpha_i + \lambda_t\beta_i + u_{it} \quad (5)$$

where α_i is the random intercept for case i, β_i is the random slope for case i, and u_{it} is the time and case-specific disturbance. The LCM uses a distinct intercept and slope to describe the linear path of the scores over time. For the intercept factor, all loadings are fixed to values of one because this latent factor equally influences all repeated measures. For the slope factor, the loadings are all spaced equally to represent a uniform time between all the repeated measurements, such that $\lambda_t = 0,1,2,3,4$.

Finally, the model uses a mean intercept and mean slope and corresponding disturbance scores to express the distribution of the intercepts and slopes. The intercept and slope equations are

$$\alpha_i = \mu_\alpha + \zeta_{\alpha i} \quad (6)$$

$$\beta_i = \mu_\beta + \zeta_{\beta i} \quad (7)$$

where μ_α and μ_β are the mean intercept and mean slope across all cases (i.e., the latent variable means), and $\zeta_{\alpha i}$ and $\zeta_{\beta i}$ are residuals. The residuals at all levels are assumed to be normally distributed with variances and a covariance between the two.

In summary, the higher-level growth factors, α and β, were connected to the latent scales (lower-level factors) with intercept factor (α) loadings fixed at one and slope factor (β) loadings set at linear time scores (i.e., 0, 1, 2, 3, and 4). The loadings of the measurement model between the factors and items were selected to produce the desired level of inter-item correlation (discussed below). After fully specifying all population model parameters, we computed the model-implied covariance matrix of the indicators using standard expressions found in structural equation modeling texts (Bollen, 1989; Bollen & Curran, 2005). Based on the item mean vector and the indicator covariance matrix, we created multivariate normal datasets by using the "MASS" package in R (Ripley et al., 2013).

Data generation and analysis procedure

We used R (R Core Team, 2018) to generate 2,000 data sets with continuous variables for each of the 36 between-subjects design cells. Because this is the first study to examine item-level imputation where entire questionnaires are missing, we opted to leave all indicators continuous for the primary simulation. We also report auxiliary simulation results in the supplement that consider binary indicators. We sketch the process of creating the parameter values here, and Table S1 in the supplemental material shows the parameter values for all unique conditions. The higher-order growth model represented the lower-order factor means as a function of the intercept and slope factor latent means and the covariance matrices of the lower-order factor were computed as a function of the growth model parameters. The parameter values used in the latent growth model were as follows: mean intercept = 20, mean slope = 0.5, intercept variance = 8, slope variance = 0.5, and the intercept-slope covariance = 0.15, and the latent disturbance terms were also fixed at 8. Importantly, these parameters were constant across all conditions of the study and did not change as the number of items increased. The mean structure parameters were selected because they imply a mean difference between the first and last wave that is commensurate with Cohen's medium effect size benchmark (when gauged relative to the baseline standard deviation). The slope variance was selected after using computer simulations to perform power analyses; a variance of .50 gave at least .80 power across most of the conditions examined. The variances of the growth factors and the disturbance variances were chosen such that latent variable at the first assessment had 50% of its variance explained by the growth factors (i.e., its intraclass correlation was .50; intercept variance/(intercept variance + disturbance variance = 8/16 = .50). In the multilevel modeling context, this intraclass correlation value is cited as typical for longitudinal designs with repeated measurements at level-1 nested within persons at level-2 (Hedges & Hedberg, 2007; Spybrook et al., 2011).

Turning to the measurement model parameters, the loading and residual variances were determined under the constraint that the item-level indicators had total variance equal to 1. With this constraint imposed, we were able to solve for the loading and residual variance values that produced within-scale item correlations equal to .3 or .5 (this was one of the manipulated factors). The measurement model parameters must be viewed in light of the constraints on growth factors (the parameters of which were constant across all conditions), the constraints on the total variance of the indicators, and the desired within-scale item correlation. The factor loadings used in this simulation (which themselves were a function of the desired inter-item correlation) imply reliabilities for sum score scales that range from 0.68 to 0.94, depending on the number of items and inter-item correlations. We believe these values represent a range of internal consistency reliability values that researchers would deem minimally acceptable in practice. Please refer to Tables

S2 in the supplementary material to see the full set of coefficient alpha reliability values for the study conditions.

After specifying all model parameters, we computed the item-level model-implied mean vector and covariance matrix. These summary quantities served as the basis for generating artificial item responses. We used the mvrnorm function from the MASS package (Ripley et al., 2013) to generate random normal item-level data. We then summed the items for case i at each time t to compute the observed scale scores. Next, we deleted all item scores from waves 3, 4, or 5 from the data simulating an MAR mechanism where the probability of missing data was positively related to the observed scale score at the first wave. Specifically, we used the first wave scale score as the cause of missingness to ensure that the MAR mechanism was satisfied for both imputation methods (i.e., scale-level missing data handling would not use the item scores, so an MAR process based on individual items might introduce biases because the mechanism is not MAR). The deletion procedure worked as follows. Using a latent variable formulation for logistic regression (Agresti, 2012), we derived the intercept and slope coefficients from a logistic model that produced an R^2 of approximately .40 between the cause of missingness (the first wave's scale score) and the underlying latent missingness probabilities for later waves (we chose .40 to ensure a relatively strong selection mechanism) (McKelvey & Zavoina, 1975). After determining the appropriate intercept and slope coefficients, we used a logistic regression equation to generate an N-row vector of missingness probabilities for waves three, four, and five. We created three different missing probabilities, with the amount of missing data increasing over time. For the low rate of missing data condition, wave 3 had a 5% missingness probability, wave 4 had a 10%, and wave 5 had a 20% missingness probability. For the high rate missing data condition, wave 3 had a 10% missingness probability, wave 4 had a 20%, and wave 5 had a 40% missingness probability. Finally, we sampled a missing data indicator for wave 3, 4, and 5 (0 = observed, 1 = missing) from a binomial distribution with success rate equal to the individual's missingness probability from the logistic regression model. All items were deleted for the wave if the observation's indicator equaled one. This deletion process produced variation in the missing data patterns, with some observations characterized by attrition (i.e., a monotone missing data pattern) and others with intermittent missingness.

We used version 2.0 of the Blimp application (Keller & Enders, 2020) to implement FCS, which is identical to the algorithm outlined by Van Buuren et al. (2006) and implemented in the popular MICE package in R (Van Buuren & Groothuis-Oudshoorn, 2011). We generated 20 imputed data sets per replication from an MCMC algorithm with 3000 burn-in iterations and 2000 thinning iterations. We chose these intervals after examining potential scale reduction factors (Gelman & Rubin, 1992) from several artificial data sets. For item-level imputation, only items were included, such that the number of variables in imputation process equaled the total number of items across all waves. Following item-level imputation, we computed a scale score by summing the items within each scale. The scale scores were the focus of the subsequent analysis phase. For scale-level imputation only scales are included in imputation and the number of variables in imputation process equaled the number of waves. Again, the scale scores are the target of the analysis phase.

In the analysis phase, we computed growth model estimates based on the observed scale scores because this is presumably how most researchers would analyze these data (i.e., a standard linear curve model, not the second-order growth model used for data generation). The factor means, variances, and covariance shown in Table S1 are also the true values for the growth model estimates of the analysis model. The estimates were computed by using full information maximum likelihood (FIML) estimation Mplus 8 (Muthén & Muthén, 1998–2017). Note that FIML estimation at this stage is not treating missing data; rather, estimation is simply based on the raw imputed data. We wrote a custom R program to pool estimates and standard errors from the 20 data sets into a single set of values. All simulation code is available upon request.

Evaluation criteria

The two criteria we used to evaluate the performance of each parameter estimate were bias and mean square error (MSE). These evaluation measures were calculated by comparing the estimates from the simulated samples to the true parameter values, which we determined analytically by transforming the second-order growth model (i.e., data-generating) parameters. Bias was expressed as the difference between an average estimate and the true value divided by the true value, and then multiplied by 100 to create a percentage (i.e., bias as a percentage of the true value). Although cutoff values are inherently subjective, authors regularly suggest that relative bias values less than 10% in absolute value are acceptable (Finch et al., 1997; Kaplan, 1988).

The second criteria we used in the study were MSE

$$MSE = \frac{1}{2000}\sum^{(\hat{\theta}-\theta)^2} \tag{8}$$

where $\widehat{\theta}$ is the parameter estimate from a particular replication within a given design cell, θ is the population parameter, and 2000 is the number of replications within a given design cell. The MSE is a composite measure that captures the accuracy and precision of an estimator, as it equals the squared bias plus the sampling variance of the parameter estimate. Because there is no reason to expect systematic biases, MSE is the outcome most likely to show differences among the methods, especially as the rate of item-level missing data and the number of items on a scale increase. To facilitate interpretation, we used MSE ratios to evaluate whether item-level imputation increases precision relative to scale-level approach imputation. We computed these MSE ratios by dividing the MSE from scale-level imputation by the MSE from item-level imputation (e.g., $MSE_{scale} \div MSE_{item}$). If estimates are unbiased, then a value greater than unity indicates that item-level imputation provided greater precision (i.e., lower sampling variation), whereas values below one mean that scale-level imputation had greater precision in comparison to item-level imputation.

Results

Bias

For all fixed and random parameter estimates, item- and scale-level imputation schemes generally produced trivial biases, although item-level imputation produced biased estimates in

a few conditions. We briefly summarize bias before moving to the MSE results, and full trellis plots of the bias values broken down by condition are found in the online supplement (see Figures S1 to S2). Averaged across all design cells, the scale-level imputation standardized percentage bias for the key parameters were as follows, mean intercept = −0.005%, mean slope = −0.007%, intercept variance = −0.495%, slope variance = 4.805%, and intercept-slope covariance = 0.023%. The item-level imputation standardized percentage bias for the key parameters included, mean intercept = −0.007%, mean slope = −0.018%, intercept variance = −0.312%, slope variance = 2.485%, and intercept-slope covariance = 4.897%.

In the majority of design cells, imputation method had no discernable impact on bias. The exception occurred in the 15 items per scale and the lowest sample size condition of $N = 250$, where item-level imputation produced negatively biased estimates of the covariance parameter; standardized bias values were −24.938% and −16.976% for the high and low missing data rate conditions, respectively. To provide further context, the average raw covariance estimate associated with the largest bias value was 0.112 (average standard error = 0.42), and the corresponding population parameter value was 0.15. Thus, the large standardized bias value was commensurate with approximately 10% of a standard error unit. We suspect that most researchers would not view this discrepancy as practically significant.

Mean square error

Since there was effectively no bias in the vast majority of cases, MSE largely captures sampling variation or precision. As explained previously, we computed MSE ratios by dividing the MSE from scale-level by the MSE from item-level imputation, such that values greater than 1 indicate that using item-level imputation has a smaller sampling variance and better overall precision. MSE ratios are useful for assessing the difference in sample size needed to equate the sampling variances of the two methods being compared. For example, an MSE ratio of 1.50 roughly indicates that the sample size for an analysis that uses scale-level imputation would need to be increased by 50% to yield the same sampling variance as item-level imputation (i.e., a 50% increase in the N should equate squared standard errors from the two procedures).

To set the stage for the MSE results, we conducted an additional simulation where item responses were independently deleted within each scale, such that some of the items within a scale were always complete. This is the scenario investigated by earlier studies that recommend item-level imputation over scale-imputation (Eekhout et al., 2014, 2015; Gottschall et al., 2012; Mazza et al., 2015; Peyre et al., 2011). The simulation procedure was identical to that described in the Methods, and the only difference was that each item – rather than an entire questionnaire – had a missing data indicator. For example, in the high missing data rate condition, wave 3 had a 10% missingness probability. In our primary simulation, entire questionnaires were removed at this rate, whereas in this supplemental simulation, each item within the questionnaire was removed at this rate. Again, this essentially created a simulation that replicated previous studies, albeit with our data generation and analysis model.

Figure 1 gives trellis plots displaying MSE ratios for all conditions subset by the strong (i.e., .50) item correlation condition. The MSE values were uniformly greater than 1, meaning that item-level imputation produced estimates that were more precise (i.e., had less sampling variation). In fact, the MSE ratios were even more dramatic than those reported in previous studies. As an example, MSE ratio greater than 2 were quite common in the plot, and this can be interpreted to mean that the sample size would need to double in order for scale-level imputation to achieve the same precision as item-level imputation. These results are perfectly consistent with previous studies, but they also serve as a stark contrast to the precision that results from entire questionnaires that are missing (our main focus). Figures S3-S6 in the online supplement show the trellis plots for bias values and MSE ratios for this simulation.

Turning to our main focus – entire waves missing – Figure 2 gives trellis plots displaying MSE ratios for all conditions subset by the moderate (i.e., .30) item correlation condition, while Figure 3 gives trellis plots displaying the conditions that are subset by the strong (i.e., .50) item correlation condition. The trellis plots show parameter-specific MSE ratios for all combinations of sample size, number of questionnaire items, and missing data rate. The first column of both figures reflects five items per scale, while the second column shows ratios for the 15 items per scale condition. All plots in both Figures 2 and 3 illustrate how sample size, number of items per scale, and missing data rate affect MSE values.

In general, the mean parameters in Figures 2 and 3 exhibit very similar patterns, and the inter-item correlation had a minimal impact on the results; the mean parameters that were substantially different between imputation approaches in the moderate correlation condition (Figure 2) were also different in the strong correlation conditions (Figure 3). For brevity, we focus on Figure 3 and will only discuss Figure 2 if there are any specific differences in results between the figures. The results for the mean intercept parameter in Figure 3 were mostly uniform, with almost all MSE ratios close to 1. The mean slope estimates in contrast had slightly more variation. In the low missing data condition (5%, 10%, and 20% at the final three waves), when the number of items-per-scale was large and the sample size was 250, the MSE ratio was smaller than 1, with scale-level imputation having 5% lower MSE than item-level imputation (first row panel). In the second-row panel, increasing the questionnaire length had no impact on MSE at sample size of 500. The third row of Figure 3 describes the high missing data condition (10%, 20%, and 40% at the final three waves) and a sample size of 250. For the condition with five items per scale, item-level imputation was equally as precise as scale-level imputation. In the same third row, but now for the condition with 15 items per scale, the MSE ratios were close to .95, with scale-level imputation showing a modest 5% lower MSE in comparison to item-level imputation. The fourth row which describes high missing data rate and a sample size of 500, increasing the questionnaire length had no impact on MSE at sample size of 500.

Considering only the intercept and slope means, the results in Figures 2 and 3 demonstrate that, for the majority of conditions there was not a substantive difference in precision between item- and scale-level imputation. There was only

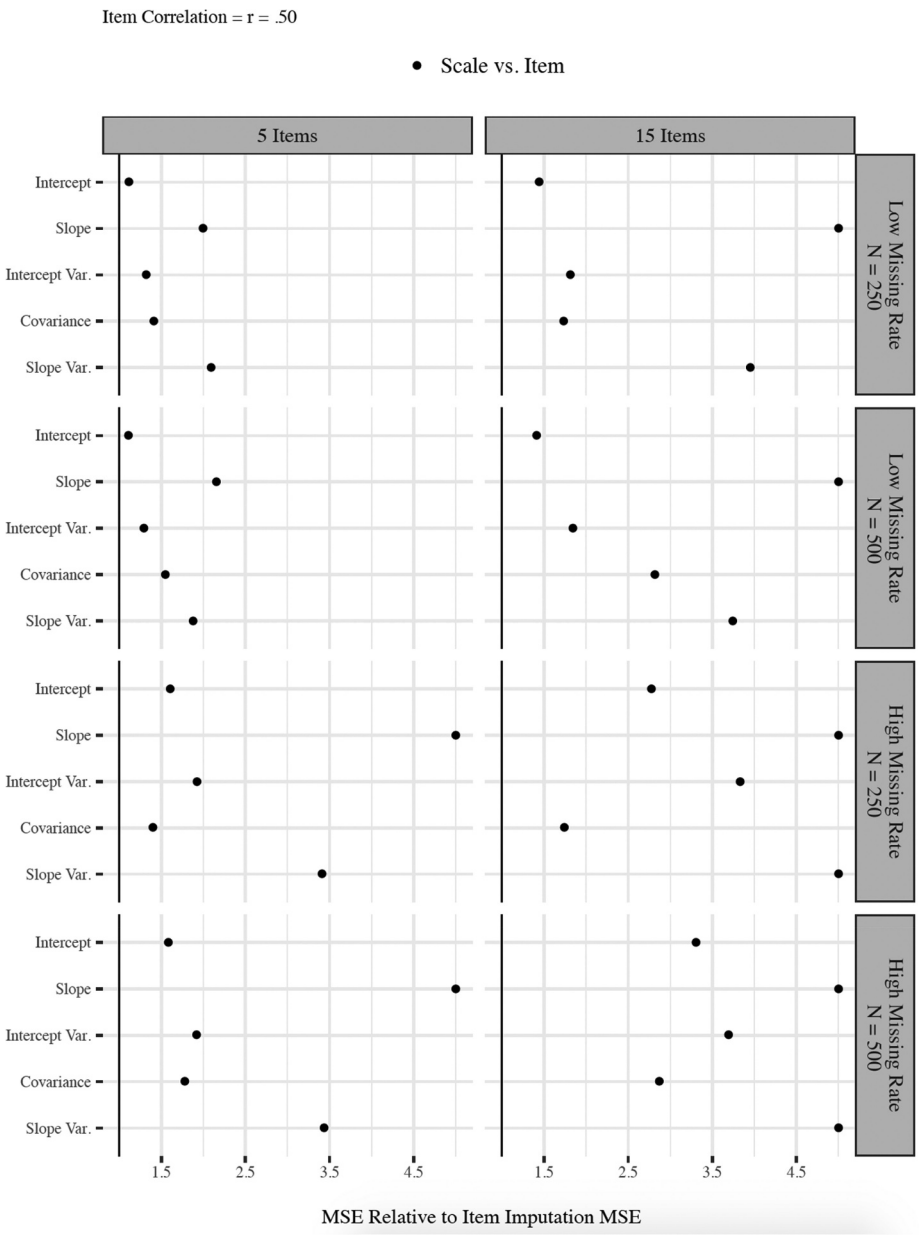

Figure 1. Intermittent missing data pattern simulation, MSE ratio values from the strong item correlation simulation featuring a two-way interaction between sample size and missing data rate with either a 5-item scale or 15-item scale.

a 5% difference in MSE values when the sample size is low and number of items is large (15), where item-level imputation produced higher MSE values than scale-level imputation, with varying differences depending on the item correlation and missing data rate. The combination of conditions involving lengthy scales, low missing data rates, and small sample sizes created the largest differences between imputation methods.

Moving our attention to the variance parameters, when the rate of missing data was low (first and second row panels) the MSE ratios of Figures 2 and 3 were close to 1 for both intercept variance and covariance. In the high missing data rate condition (third and fourth row panels) for both the moderate and strong item correlation, when the questionnaire length is 15 items-per-scale, the moderate item correlation (Figure 2) scale level imputation had close to a 95% as large MSE as item-level for the covariance parameter, while the MSE ratio for the intercept variance remained close to 1. For the strong item correlation (Figure 3), when the sample size is 250 and number of items is large (third row panel, second column) the results show that scale level imputation had a smaller MSE value than item-level imputation by close to 11% for covariance and 5% for the intercept variance, which means that the scale level imputation was more precise than item-level imputation, when the missing data rate was high, the number of items-per-scale was large, and the sample size was small.

The results for the slope variance parameter indicated larger differences in MSE values between item-level and scale-level imputation. When missing data rate is high and the number of items large (15), scale-level imputation had a substantial benefit on precision in comparison to item-level imputation for the slope variance parameter, and there is no situation in which

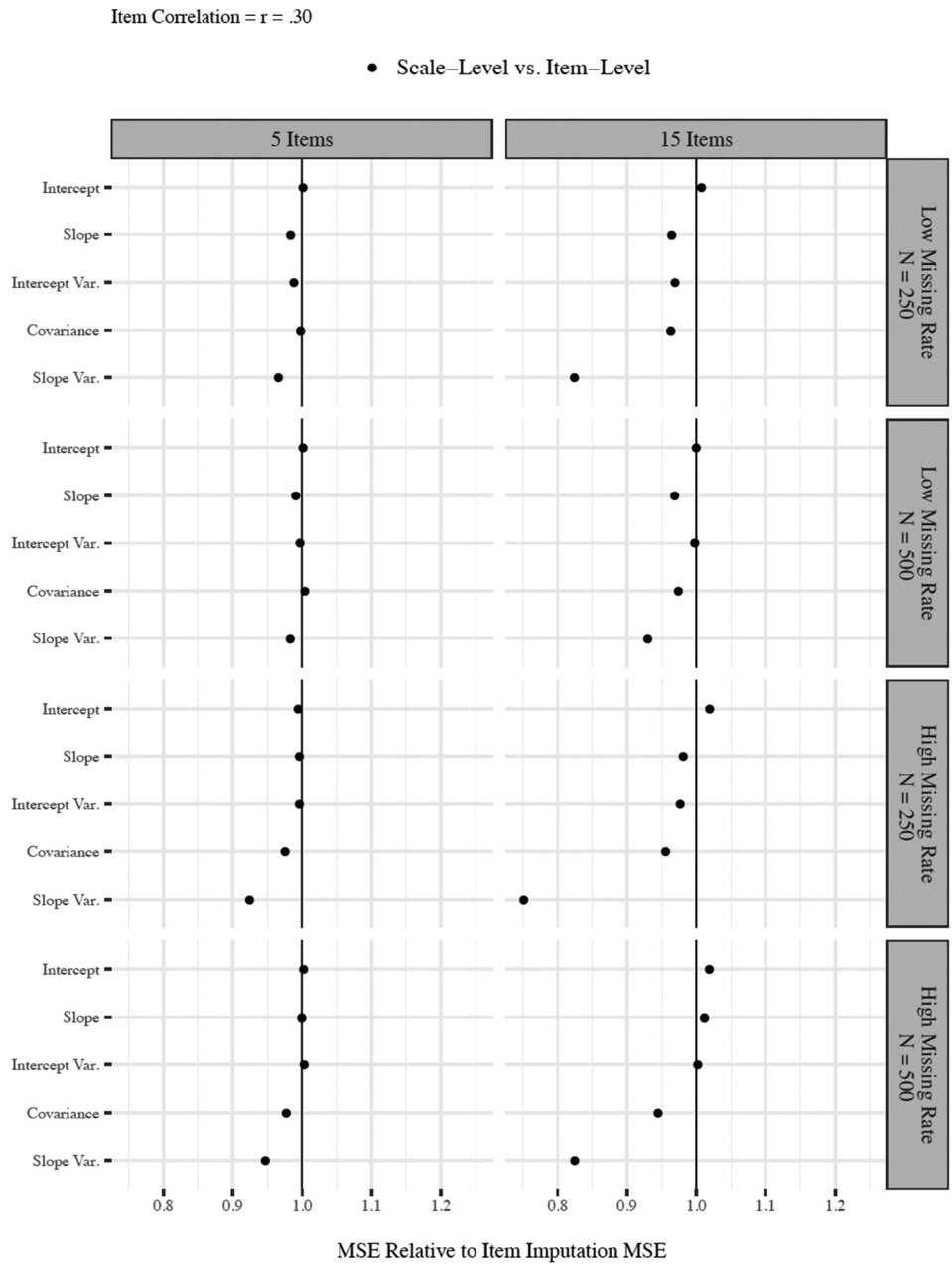

Figure 2. Primary simulation, MSE ratio values from the moderate item correlation simulation featuring a two-way interaction between sample size and missing data rate with either a 5-item scale or 15-item scale.

item-level imputation produced a more precise estimate of this parameter. In the first row of Figure 2, when the number of items per scale increased to 15, the MSE ratios for slope variance were substantially less than 1. The scale-level imputation MSE was 85% as large as that of item-level imputation MSE, meaning that scale-level provided greater precision or less sampling variation. However, when we increased the sample size to 500, as shown in the second row of Figure 2, the MSE ratio for 5 item-per-scale remained close to 1 and the MSE ratio for the 15 item-per-scale was .93. This makes sense, as increasing the sample size should mitigate a noisy estimate that results from imputing many item responses. In the third row, for the five-item per scale condition, the MSE ratios were again smaller than 1, showing that scale-level imputation had a lower MSE value compared to item-level by close to 10%. The 15 item per scale condition in the third row of Figure 2 produced an MSE ratio of .46 (it is fixed at the minimum of the horizontal axis, .75), which means that scale-level was substantially more precise than item-level imputation. The effect of increasing the number of items from 5 to 15 was less pronounced for the conditions of high missing data rate and a sample size of 500 (fourth row). Again, the trends in Figure 2 were not meaningfully different from those in Figure 3, so no further discussion is warranted.

Additional simulations

The previous simulation results addressed our main question, which was whether the recommendation to use item-level imputation is appropriate for situations where entire

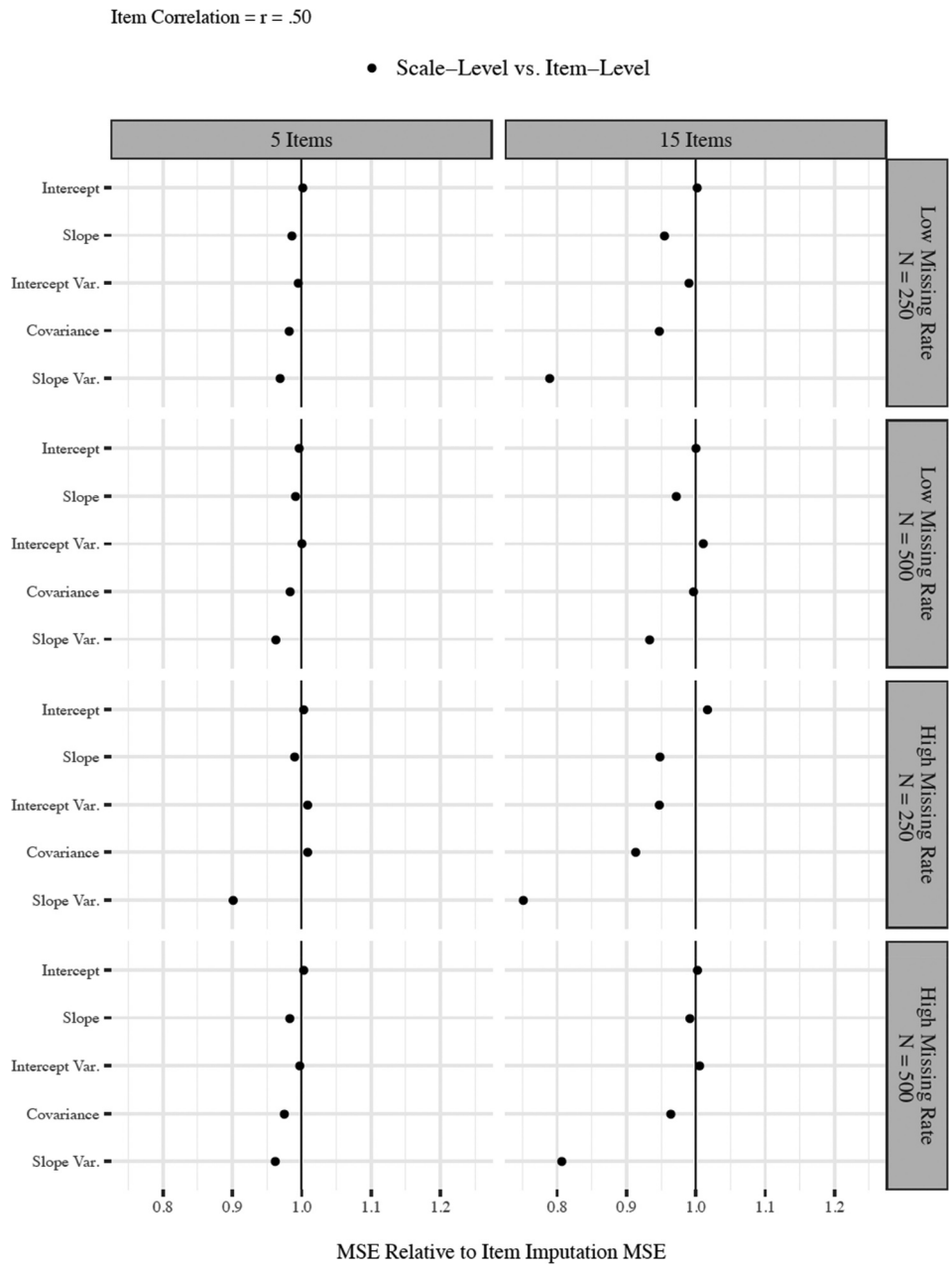

Figure 3. Primary simulation, MSE ratio values from the strong item correlation simulation featuring a two-way interaction between sample size and missing data rate with either a 5-item scale or 15-item scale.

questionnaires are missing because respondents skipped one or more data collection waves; as mentioned previously, published studies have focused exclusively on the situation where some but not all of the items within a scale are missing. The simulation results clearly show that the benefits of item-level imputation essentially vanish in this situation. At best, item-level imputation was equivalent to scale-level imputation in terms of precision (e.g., the growth factor means), but the slope variance estimate from this approach could be quite noisy, to the point where the sampling variance was nearly double that of scale-level imputation. Thus, the conclusion is that there is very little reason to invoke an item-level imputation scheme to the missing data pattern we investigated.

We completed two additional simulations that attempt to generalize our main results to a broader range of conditions. The first additional simulation was the same as our primary simulation but with a more complex missingness process. You may recall that our primary simulation used a strategy where the missingness probabilities increased as a function of the wave 1 scale scores. Following Collins et al. (2001), we investigated a "convex" missing at random process where the missingness probabilities increase as the wave 1 scale scores increase *or* decrease. The overall missingness rates remained the same as those from the primary study.

The convex missing data pattern simulation produced trivial biases for all fixed parameter estimates and for most of the variance parameters, as you might expect. Similar to the

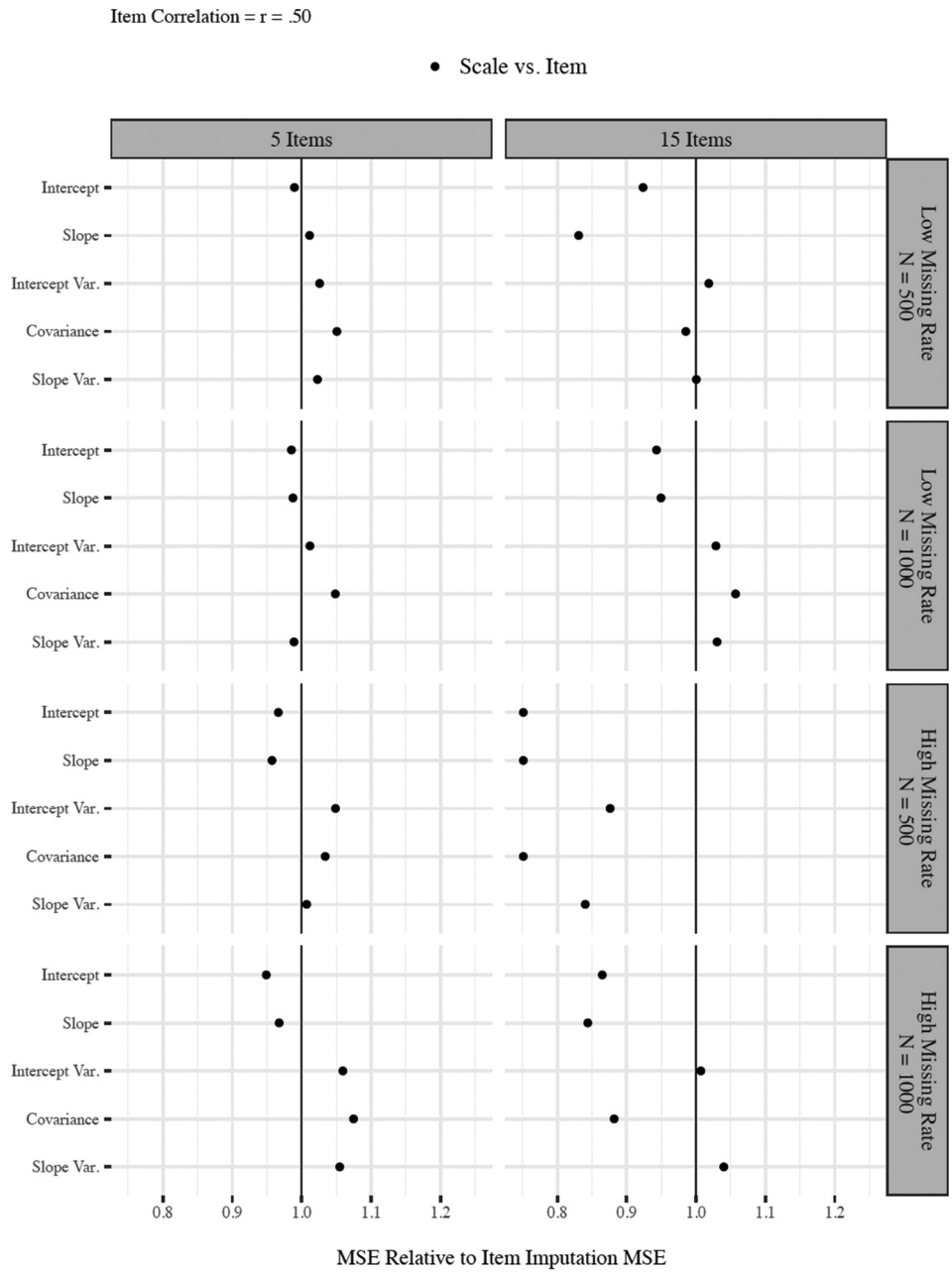

Figure 4. Binary items simulation, MSE ratio values from the strong item correlation simulation featuring a two-way interaction between sample size and missing data rate with either a 5-item scale or 15-item scale.

primary simulation, the exception occurred in the 15 items per scale and the lowest sample size condition of $N = 250$, where item-level imputation produced negatively biased estimates of the covariance parameter (Figures S7 and S8) . Using a convex missing data pattern did not alter the original results from the main simulation. Scale-level imputation still produced lower MSE values for slope variance estimate when the sample size was small and the number of items was large; this was true for both high and low missing data rates. Figures S9 and S10 in the online supplement show the trellis plots for this simulation.

Next, our main simulation focused on continuous indicators or items. We chose this feature to investigate an optimal scenario where linear regression imputation models are appropriate and to provide comparability to published studies. Of course, researchers are most likely to encounter categorical item responses that are ultimately summed to create a quasi-interval scale for later analysis. To investigate whether our results generalize to discrete item responses, we conducted a second additional simulation that used binary indicators. There were only main changes to the simulation procedure described in the Methods section. First, the set of conditions of high missing data rate (10%, 20%, and 40% missingness probability), sample size condition of 250, and 15 items per scale did not converge when using item-level imputation. Convergence issues are not uncommon when attempting to impute a large number of categorical variables with a small sample size, and this result has been reported elsewhere in the literature (Rombach et al., 2018; Simons et al., 2015). Therefore, we only used the conditions with sample size of 500 and 1000 in the plots for bias values and MSE ratios for this simulation

(Figures S11-S14 in the online supplement). Second, we dichotomized the continuous indicators by assigning an item response equal to 0 if the continuous indicator was below 0 (the mean of the wave 1 items) and equal to 1 if the continuous score was greater than 0. By applying the same categorization threshold to all items, we created a scenario where the proportion of 1 s linearly increased over time (because the continuous item means were 0 at wave 1 and increased over time). The online supplement provides a detailed description of this simulation.

In the simulation with binary items, for all fixed and random parameter estimates, item- and scale-level imputation scheme generally produced trivial biases. The only exception was for the estimates of the covariance parameter under high missing data rate, especially at large number of items per scale (15) and the smaller sample size condition (500). The bias values for those set of conditions where 21% and 14% for the moderate and strong inter-item correlations, respectively (see Figures S11 and S12). Moving our attention to Figure 4, which displays the MSE ratios for all conditions subset by the strong (i.e., .50) item correlation condition; the results for the mean intercept parameter in Figure 4 were mostly uniform, with almost all MSE ratios between .92 and 1. The exception occurred in the conditions pairing a high missing data rate and 15 items per scale, where the MSE ratios were in the .70 to .80 range (i.e., scale-level imputation was more precise; see the second column of the third and fourth row panel). Moving our attention to the mean slope estimates, most of the MSE ratios, with the exceptions mentioned below, ranged from .91 to 1.07 which does not show a substantial difference in precision between imputation methods. The exceptions were for the conditions of low missing data rate, 15 items per scale, and sample size of 500 (first row panel, second column), as well as the high missing data rate, 15 items per scale, and both sample size (third and fourth row panel, second column), where scale-level imputation was substantially more precise than item-level imputation.

The MSE ratios for intercept and slope variance and the covariance do not show substantial differences in precision between imputation methods at the lowest missing data rates (first and second row-panels). Scale-level imputation was more precise at estimating the intercept variance, but only in the conditions involving a high missing data rate, 15 items per scale, and sample size equal to 500; for all other conditions, precision differences were relatively small. For the slope variance parameter, the conditions involving a high missing data rate, 15 items per scale, and sample size of 500 showed an advantage for scale-level, where the MSE ratio was .76, indicating that scale-level MSE was about three-quarters the size of the item-level MSE. Scale-level imputation also show advantage in precision in estimating the covariance parameter at the high missing data condition and 15 items per scale (third and fourth row panels, and the second column). The MSE ratios were .56 and .88 for the 500 and 1000 sample size conditions, respectively. Similar to the primary simulation, item-level imputation did not provide meaningful gains in precision in any of the conditions. When the number of items per scale was large and the missing data rate was high, scale-level again had an advantage in precision in comparison to item-level imputation when estimating the slope variance. However, using binary data did alter the results found in the primary simulation. Scale-level imputation also showed an advantage in precision over item-level when estimating the mean intercept and slope parameters when high levels of missing data were present.

Discussion

The simulation study examined the performance of item-level and scale-level imputation in longitudinal studies where all questionnaire items are missing because a participant skipped one or more waves of data collection. In contrast to past studies where handling missing data at the item level provided substantial improvements in precision (Eekhout et al., 2015; Gottschall et al., 2012; Mazza et al., 2015), the results of the current study show that under conditions that are common to longitudinal studies, handling missing data at the scale-level is superior. As mentioned previously, the key difference between our study and previous work is the missing data pattern; we investigated a situation where *all* questionnaire items are missing because of wave non-response, whereas prior work investigates the scenario were some but not all of the items are missing. These two patterns lead to very different results and conclusions.

There were no major differences in bias, and the imputation strategy had the biggest impact on precision. For the latent means (i.e., average intercept and slope), item-level imputation produced no advantages in the way of sampling variation, and item- and scale-level imputation were effectively equivalent. With a few exceptions, relatively small precision differences were also observed for other parameters, although there was virtually no situation were item-level imputation was preferred; at best it produced roughly equivalent precision. Contrary to Gottschall et al. (2012) and others, the result of the simulation showed that when the number of items per scale was large and the missing data rate was high, item-level imputation did not provide meaningful gains in precision, not even at small sample sizes. The MSE differences for slope variance were quite large in some cases, with scale-level imputation producing more precise estimates than item-level imputation. Perhaps not surprisingly, this effect was most evident in the high missing data condition compared to the low missing data rate condition. For both the high and low missing data conditions, increasing questionnaire length had a strong impact in MSE values (e.g., increasing the precision gap between the two approaches), however the size of these differences was somewhat larger in the high missing data condition. Again, there was no condition in which item-level imputation produced a more precise estimation of variance components in comparison to scale-level imputation.

We present two possible explanations why scale-level imputation performs better than item-level imputation with this particular missingness pattern. First, item-level imputation is less parsimonious than the scale-level imputation model. This is especially detrimental in longitudinal data with wave non-response, where item-level imputation would require a large number of parameters to estimate missing data for scales where all items are missing. The second explanation

is that the growth model's covariance structure actually induced fairly strong correlations between waves. One important disadvantage of scale-level imputation in cross-sectional studies is that it ignores the strongest source of correlation in the data – the inter-item correlations among items from the scale score – in favor of weaker between-scale correlations. However, the situation changes in a longitudinal framework where the correlations between scales are stronger than what is typical in cross-sectional studies. When entire waves are missing, there are no within-scale correlations to leverage, so the only source of explained variance is between measures at different waves. As a result, item- and scale-level imputation end up using the same information – between-wave correlations – although they do so with different numbers of parameters.

Limitation and conclusion

As with any simulation study, our research has limitations that are important to highlight. Using the first wave scale score as the cause of missing data in the deletion process could potentially be a limitation. In practice, the causes of missingness can be more complex, as they may vary across subjects and across variables, resulting in a range of possible missingness processes. Our follow-up simulation with a convex pattern did not produce meaningful differences, but this missingness process is still an over-simplification to what might occur in practice. Similarly, we only investigated five waves of repeated measurements. Although our results probably generalize reasonably well to studies with fewer waves of data, the model complexity associated with item-level imputation is exacerbated with additional assessments and likely requires much larger samples than we investigated here. Another important limitation is our choice of data-generating parameters, as the growth model parameters – in particular, the slope variance and the intercept-slope covariance – could have a substantial impact on the between-wave scale correlations and thus the relative precision of different missing data handling methods. We chose the intercept variance to produce an ICC that is typical of longitudinal data, and we similarly chose the slope variance based on preliminary power analyses. In a similar vein, we used a relatively simple model that did not include predictors of growth or distal outcomes. The unique features of our model may not generalize to the myriad different longitudinal models that researchers apply in practice, and further research should investigate different analytic models.

In addition, the study cannot be generalized to a strict MNAR process where missingness is due to the item responses themselves. In fact, there are many ways in which an MNAR process could occur in a longitudinal study. Among other things, item-level missingness could be due to the underlying item response, scale-level missingness could be due to one's standing on the latent factor or scale, and item- or scale-level missingness could be due to the growth factors (e.g., one's trajectory determines whether later responses are missing). MNAR processes could also result from applying scale-level imputation to an MAR process where items are the cause of missingness. The impact of such a process would surely depend on whether items or scales were the cause of missingness, the proportion of NMAR items, the strength of the selection mechanism, among other things. The broad nature of this process is beyond the scope of this paper, but it is an important avenue for future research.

In sum, we investigated performance of item-level and scale-level imputation in longitudinal studies where all questionnaire items are missing because a participant skipped one or more waves of data collection. Contrary to suggestions from the literature the result of the simulations showed that item-level imputation did not provide meaningful gains in precision. If participants skip an entire wave, our results suggest that imputation at the scale level may provide more precise estimates of growth model parameters compared to the imputation at the item-level, specially when the number of items per scale is large and the missing data rate are high. Thus, the conclusion is that there is very little reason to invoke an item-level imputation scheme to the missing data pattern we investigated.

Funding

This work was supported by Institute of Educational Sciences award [R305D150056].

References

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

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling* (pp. 243–277). Lawrence Erlbaum Associates.

Bollen, K. A. (1989). *Structural equations with latent variables* (Vol. 210). John Wiley & Sons.

Bollen, K. A., & Curran, P. J. (2005). *Latent curve models* (Vol. 467). John Wiley & Sons, Inc.

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. https://doi.org/10.1037/1082-989X.6.4.330

Eekhout, I., de Vet, H. C., Twisk, J. W., Brand, J. P., de Boer, M. R., & Heymans, M. W. (2014). Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. *Journal of Clinical Epidemiology*, *67*, 335–342. https://doi.org/10.1016/j.jclinepi.2013.09.009

Eekhout, I., Enders, C. K., Twisk, J. W. R., de Boer, M. R., de Vet, H. C. W., & Heymans, M. W. (2015). Analyzing incomplete item scores in longitudinal data by including item score information as auxiliary variables. *Structural Equation Modeling-a Multidisciplinary Journal*, *22*, 588–602. https://doi.org/10.1080/10705511.2014.937670

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

Finch, J. F., West, S. G., & MacKinnon, D. P. (1997). Effects of sample size and nonnormality on the estimation of mediated effects in latent variable models. *Structural Equation Modeling-a Multidisciplinary Journal*, *4*, 87–107. https://doi.org/10.1080/10705519709540063

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

Ginkel, J. R., Ark, L. A., & Sijtsma, K. (2007). Multiple imputation for item scores when test data are factorially complex. *British Journal of Mathematical and Statistical Psychology*, *60*, 315–337. https://doi.org/10.1348/000711006x117574

Gottschall, A. C., 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. https://doi.org/10.1080/00273171.2012.640589

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling*, *10*, 80–100. https://doi.org/10.1207/S15328007sem1001_4

Grimm, K. J., Ram, N., & Estabrook, R. (2016). *Growth modeling: Structural equation and multilevel modeling approaches*. Guilford Publications.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*, 60–87. https://doi.org/10.3102/0162373707299706

Hughes, R. A., White, I. R., Seaman, S. R., Carpenter, J. R., Tilling, K., & Sterne, J. A. C. (2014). Joint modelling rationale for chained equations. *BMC Medical Research Methodology*, *14*, 28. https://doi.org/10.1186/1471-2288-14-28

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*, 183–202. https://doi.org/10.1007/bf02289343

Kaplan, D. (1988). The impact of specification error on the estimation, testing, and improvement of structural equation models. *Multivariate Behavioral Research*, *23*, 69–86. https://doi.org/10.1207/s15327906mbr2301_4

Keller, B. T., & Enders, C. K. (2020). *Blimp user's guide (version 2.2)*. Los Angeles.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford press.

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. https://doi.org/10.1080/00273171.2015.1068157

Mcardle, J. J., & Epstein, D. (1987). Latent growth-curves within developmental structural equation models. *Child Development*, *58*, 110–133. https://doi.org/10.1111/j.1467-8624.1987.tb03494.x

McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *The Journal of Mathematical Sociology*, *4*, 103–120. https://doi.org/10.1080/0022250X.1975.9989847

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, *55*, 107–122. https://doi.org/10.1007/BF02294746

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Newsom, J. T. (2015). *Longitudinal structural equation modeling: A comprehensive introduction*. Routledge.

Peyre, H., Leplège, A., & Coste, J. (2011). Missing data methods for dealing with missing items in quality of life questionnaires. A comparison by simulation of personal mean score, full information maximum likelihood, multiple imputation, and hot deck techniques applied to the SF-36 in the French 2003 decennial health survey. *Quality of Life Research*, *20*, 287–300.

R Core Team. (2018). *R: A language and environment for statistical computing*.

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., Firth, D., & Ripley, M. B. (2013). *Package 'mass'*. Cran R, 538.

Rombach, I., Gray, A. M., Jenkinson, C., Murray, D. W., & Rivero-Arias, O. (2018). Multiple imputation for patient reported outcome measures in randomised controlled trials: Advantages and disadvantages of imputing at the item, subscale or composite score level. *BMC Medical Research Methodology*, *18*, 87. https://doi.org/10.1186/s12874-018-0542-6

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

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

Schafer, J. L. (2001). Multiple imputation with PAN. In A. G. Sayer & L. M. Collins (Eds.), *New methods for the analysis of change* (pp. 355-377). American Psychological Association.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177. https://doi.org/10.1037//1082-989x.7.2.147

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. https://doi.org/10.1207/s15327906mbr3304_5

Simons, C. L., Rivero-Arias, O., Yu, L.-M., & Simon, J. (2015). Multiple imputation to deal with missing EQ-5D-3L data: Should we impute individual domains or the actual index? *Quality of Life Research*, *24*, 805–815. https://doi.org/10.1007/s11136-014-0837-y

Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., & Raudenbush, S. W. (2011). *Optimal design plus empirical evidence: Documentation for the "optimal design" software*. William T. Grant Foundation. Retrieved on November, 5, 2012.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, *16*, 219–242. https://doi.org/10.1177/0962280206074463

Van Buuren, S. (2012). *Flexible imputation of missing data (Chapman & Hall/CRC interdisciplinary statistics)*. Chapman and Hall/CRC.

Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*, 1049–1064. https://doi.org/10.1080/10629360600810434

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*, 1–67. https://doi.org/10.18637/jss.v045.i03

Van Ginkel, J. R., Sijtsma, K., Van der Ark, L. A., & Vermunt, J. K. (2010). Incidence of missing item scores in personality measurement, and simple item-score imputation. *Methodology*, *6*, 17–30. https://doi.org/10.1027/1614-2241/a000003

Vermunt, J. K., van Ginkel, J. R., van der Ark, L. A., & Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology*, *38*, 369–397. https://doi.org/10.1111/j.1467-9531.2008.00202.x

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. https://doi.org/10.1177/0049124112460373