Multivariate Behavioral Research

ISSN: 0027-3171 (Print) 1532-7906 (Online) Journal homepage: http://www.tandfonline.com/loi/hmbr20

A Comparison of Item-Level and Scale-Level Multiple Imputation for Questionnaire Batteries

Amanda C. Gottschall , Stephen G. West & Craig K. Enders

To cite this article: Amanda C. Gottschall , Stephen G. West & Craig K. Enders (2012) A Comparison of Item-Level and Scale-Level Multiple Imputation for Questionnaire Batteries, Multivariate Behavioral Research, 47:1, 1-25, DOI: 10.1080/00273171.2012.640589

To link to this article: http://dx.doi.org/10.1080/00273171.2012.640589

Multivariate Behavioral Research, 47:1–25, 2012

ISSN: 0027-3171 print/1532-7906 online
DOI: 10.1080/00273171.2012.640589

Psychology Press
Taylor & Francis Group

A Comparison of Item-Level and Scale-Level Multiple Imputation for Questionnaire Batteries

Amanda C. Gottschall, Stephen G. West, and Craig K. Enders
Arizona State University

Behavioral science researchers routinely use scale scores that sum or average a set of questionnaire items to address their substantive questions. A researcher applying multiple imputation to incomplete questionnaire data can either impute the incomplete items prior to computing scale scores or impute the scale scores directly from other scale scores. This study used a Monte Carlo simulation to assess the impact of imputation method on the bias and efficiency of scale-level parameter estimates, including scale score means, between-scale correlations, and regression coefficients. Although the choice of imputation approach had no influence on the bias of scale-level parameter estimates, it had a substantial impact on efficiency, such that item-level imputation consistently produced a meaningful power advantage. The simulation results clearly supported the use of item-level imputation. To illustrate the differences between item- and scale-level imputation, we examined predictors of 7th-grade academic self-efficacy in a sample of 595 low-income Mexican Origin adolescents in a planned missingness design. The results of the empirical data analysis were consistent with those of the simulation and also suggested that researchers should be cautious when implementing planned missing data designs that necessitate scale-level imputation.

Behavioral science researchers routinely administer questionnaires comprised of several items that measure the same underlying construct. For example, the Beck

Correspondence concerning this article should be addressed to Amanda C. Gottschall, Parenting & Family Research Center, University of South Carolina, 1233 Washington Street, 2nd Floor, Columbia, SC 29208. E-mail: Amanda.Gottschall@sc.edu

Depression Inventory (Beck, Ward, & Mendelson, 1961) includes 21 items that measure different aspects of depression (e.g., sadness, loss of energy, changes in appetite). Although the item-level responses are often the focus of psychometric analyses, researchers typically use scale scores that sum or average the items to address their substantive questions. Item-level missing values are a common feature of questionnaire data. Among other reasons, item responses may be missing because a participant overlooks or refuses certain items, because the respondent fails to complete a timed test, or because a researcher chooses to administer a subset of questionnaire items to each respondent (e.g., a planned missing data design; Graham, Taylor, Olchowski, & Cumsille, 2006). Item-level missing values pose interesting challenges for scale-level analyses.

The last decade has seen a noticeable shift to missing data handling methods that assume a missing at random (MAR) mechanism, where the propensity for missing data on a variable Y is fully explained by the observed values of other variables (Rubin, 1976). As applied to questionnaire data, MAR holds when the propensity to skip a particular questionnaire item is related to other variables (e.g., demographic variables such as age or socioeconomic status or items measuring other constructs) but not to the would-be values of that item. MAR also holds when the propensity for missing data is unrelated to other variables, as in a planned missing data design (which typically involves a missing completely at random mechanism).

In the behavioral sciences, maximum likelihood estimation and multiple imputation are the principal methods for obtaining MAR-based estimates. Maximum likelihood uses the observed data to estimate model parameters, whereas multiple imputation fills in the missing values in a preliminary imputation phase. In many situations, the decision to use maximum likelihood estimation or multiple imputation is largely one of personal preference because the methods yield similar results (Collins, Schafer, & Kam, 2001; Schafer, 2003). However, maximum likelihood estimation is arguably less flexible with item-level missing data, particularly when the goal is to compute and analyze scale scores. Unless a researcher is willing to recast questionnaire items as indicators of a latent factor, a scale score must be treated as incomplete, even if an individual responds to some of the items within the scale. Multiple imputation provides a flexible mechanism for dealing with item-level missingness because it separates the missing data handling from the analysis. For example, a researcher applying multiple imputation to questionnaire data could either impute the incomplete items prior to computing scale scores or impute the scale scores directly from other scale scores; we henceforth refer to these methods as item-level and scale-level imputation, respectively.

To illustrate the differences between item- and scale-level imputation, consider a data set with two scale scores, X and Y, each of which is comprised of three items, x_1 to x_3 and y_1 to y_3, respectively. Further, suppose that a group

of participants have missing values on x_1. By far, the most common imputation approach in the social sciences is to assume a multivariate normal model for the incomplete variables (e.g., data augmentation; Schafer, 1997). Under this scheme, an iterative algorithm generates regression equations for each unique missing data pattern, whereby the incomplete variables serve as outcomes and the complete variables are predictors. For example, the item-level imputation model at a particular iteration would use the five observed items to impute the missing x_1 item score, as follows:

$$x_{1i}^{*} = [b_0 + b_1 x_{2i} + b_2 x_{3i} + b_3 y_{1i} + b_4 y_{2i} + b_5 y_{3i}] + r_i, \tag{1}$$

where x_{1i}^{*} is the imputed value for case i, the collection of terms in brackets comprise a predicted score, and r is a random residual from a normal distribution. The resulting scale score for case i is the sum or average of the imputed and observed items. In contrast, scale-level imputation ignores the item-level responses and treats X or Y as missing if one or more of the component items are incomplete. The imputation regression model for this example is

$$X_i^{*} = [b_0 + b_1 Y_i] + r_i, \tag{2}$$

where X_i^{*} is the imputed X scale score, Y_i is the observed Y scale score, and the remaining terms are the same as Equation 1.

Equations 1 and 2 illustrate that item- and scale-level imputation differ in the number and the quality of the predictor variables that they incorporate into the imputation process. Specifically, notice that the scale-level imputation model does not include the two complete items, x_2 and x_3. Rather, the approach treats the X scale score as missing and uses the observed Y scale to generate imputations. Because within-scale correlations tend to be stronger than between-scale correlations, scale-level imputation effectively excludes the best predictors of the missing scale score in favor of explanatory variables with weaker correlations. At an intuitive level, relying on weaker between-scale predictors can reduce the precision of the parameter estimates, but the magnitude of this power reduction is difficult to predict.

To date, methodologists have devoted little attention to the differences between item- and scale-level imputation. Belin, Datt, Desmond, and Ganz (1999) applied the two approaches to a real data set and found some similarities and some differences (e.g., an estimate that was significant under one method was nonsignificant under the other). However, using a single real data set does not allow one to draw definitive conclusions about accuracy and power differences. Several studies have compared the performance of item-level imputation with other missing data handling approaches (e.g., listwise deletion, mean imputation, hot deck imputation), but none included scale-level imputation as a comparison (Bernaards & Sijtsma, 1999, 2000; Enders, 2003; Gmel, 2001; McDonald,

Thurston, & Nelson, 2000; Van Ginkel, 2010; Van Ginkel, Van der Ark, & Sijtsma, 2007a, 2007b).

Studying the performance of item- and scale-level imputation is important because this issue is widely relevant to substantive researchers in the behavioral sciences. Based on the previous discussion, it might seem that item-level imputation is automatically preferred over scale-level imputation because it incorporates stronger correlates of the incomplete variable. However, this is not necessarily true. Imputing at the item level is often difficult or impossible. As an upper limit, the number of variables in a regression-based imputation model cannot exceed the number of cases, although in practice the maximum number of variables may be much smaller. In a study with several multiple-item measures (e.g., a longitudinal study where researchers administer multiple questionnaires at each wave), the item-level imputation model can get prohibitively large, leading to convergence problems. Scale-level imputation avoids this problem because the number of cases tends to far exceed the number of scale scores. If scale-level imputation produces only a modest power reduction, then researchers will have empirical support for applying this approach to large-scale imputation problems (Graham, 2009).

Comparing item- and scale-level imputation also bears on the development of planned missing data designs such as those described by Graham and colleagues (e.g., the three-form design; Graham, Hofer, & MacKinnon, 1996; Graham et al., 2006). The basic idea behind these designs is to administer a subset of the questionnaire items to each respondent (e.g., for the purposes of reducing respondent burden or data collection costs). When employing planned missing data designs, researchers can assign a scale's items to the same test form (e.g., participants are administered some questionnaires but not others) or spread the items across different test forms (e.g., participants are administered a subset of items from all questionnaires). Although logistical concerns often dictate this choice (e.g., see Graham et al., 2006, for a discussion), it is important to point out that grouping items together generally requires scale-level imputation, whereas spreading items across forms can accommodate item- or scale-level imputation. To date, the methodological literature has not thoroughly addressed the role of imputation on this important design choice. Consequently, examining the relative performance of item- and scale-level imputation can enhance our understanding of these increasingly popular designs.

Given the lack of existing research, the purpose of this study is to examine the relative performance of item- and scale-level imputation. To this end, we designed a Monte Carlo simulation study that allowed us assess the impact of imputation method on the bias and efficiency (i.e., power) of scale-level parameter estimates (e.g., scale score means, between-scale correlations, and regression coefficients). The next section describes the simulation studies in more detail. As an aside, we chose to limit our investigation to multiple imputation

because this method has a strong theoretical foundation and a large body of empirical literature supporting its use (e.g., for details see Enders, 2010; R. J. A. Little & Rubin, 2002; Schafer, 1997; Schafer & Graham, 2002; Schafer & Olsen, 1998; Sinharay, Stern, & Russell, 2001). However, it is important to point out that a variety of single imputation techniques can be applied to item-level missing data. Perhaps the most popular approach is to compute scale scores by averaging the available item responses (equivalent to imputing the missing items with an individual's average item response). In our view, the fact that these approaches lack a theoretical rationale is problematic because it is difficult to make general predictions about their performance. Although some of these approaches have limited empirical support (e.g., see Van Ginkel, 2010), we chose not to include them in our study.

METHOD

Manipulated Factors

We implemented a full factorial design that was comprised of one within-subjects factor (imputation approach) and six between-subjects factors: (a) rate of item-level missingness, (b) number of scales, (c) magnitude of between-scale correlations, (d) number of items per scale, (e) homogeneity of within-scale correlations, and (f) sample size. We chose the between-subjects factors because we anticipated that they would exert an influence on the relative performance of the two imputation approaches. For example, based on Equations 1 and 2, we expected the performance of item-level imputation to improve as the number of scale items increased and the homogeneity of the within-scale correlations increased (e.g., correlations were uniformly strong in magnitude). In contrast, we expected the performance of scale-level imputation to improve as the number of scales increased and the magnitude of the between-scale correlations increased.

To choose realistic levels for each manipulated factor, we reviewed published factor analytic studies that appeared in the 2006 and 2007 volumes of the American Psychological Association journal *Psychological Assessment*. Because these articles rarely reported item-level missing data rates, we determined realistic values for this factor by examining the responses to a large multiscale survey that was administered to approximately 2,000 students enrolled in an introductory undergraduate psychology course. Based on our literature review, we implemented the following levels for the between-subjects factors: (a) item-level missing data rate (5% or 15%), (b) number of scales (3 or 6), (c) magnitude of the between-scale correlations ($r = .10$ or $.50$), (d) number of items per scale (3 or 12), (e) homogeneity of within-scale correlations (uniform, moderate homogeneity, low homogeneity), and (f) sample size ($N = 200$, 400, or 800).

These conditions represent approximate lower and upper limits of the study features that were common in the published articles that we reviewed. This combination of conditions produced 144 between-subjects design cells and a total of 288 design cells.

Population Model

We used a factor analysis model as the data-generating population model for the simulation studies. To illustrate, Figure 1 shows the population model for a condition with three factors (i.e., scales) and three items per scale. Within a given design cell, the population model had identical factor correlations (either .10 or .50), had the same number of items per factor (3 or 12), and had the same pattern of loadings. To manipulate the homogeneity of the within-scale correlations, we introduced three factor loading patterns. The uniform condition had standardized factor loadings of $\lambda = .90$, such that the model-implied correlations were identical within each scale. In the moderate homogeneity condition, one third of the items had standardized factor loadings equal to $\lambda = .90$, one third had $\lambda = .75$, and one third had $\lambda = .60$. In the low homogeneity condition, one third of the standardized loadings were set at $\lambda = .90$, one third were set at $\lambda = .60$, and one third were set at $\lambda = .30$. The moderate and low homogeneity conditions produced model-implied correlations that varied in magnitude within a given scale. Finally, all uniqueness terms had a variance equal to 1 minus λ^2, such that the continuous indicators had unit variance in the population.

As seen in Figure 1, the population model also included an additional auxiliary variable that represented the cause of missingness (e.g., this variable can be thought of as a set of measured participant characteristics that impact the propensity for missing data). The auxiliary variable was always correlated with the latent factors at $r = .40$, which is roughly the minimum value for an auxiliary variable to exert a noticeable influence during imputation (Collins et al., 2001). As described in the next section, we used this variable to introduce an MAR missing data mechanism.

Data Generation

We used SAS 9.1 to generate 2,000 data sets within each of the 144 between-subjects design cells. To begin, we used the RANNOR function in PROC IML to generate random normal item-level data, and we then used Cholesky decomposition to transform the variables to be consistent with the population model-implied correlation matrix (Kaiser & Dickman, 1962). Next, we used thresholds of $z = -1.28$, -0.69, 0.69, and 1.28 to categorize the continuous items into 5-point discrete scales. This produced symmetric ordinal distributions that mimicked a Likert response format.

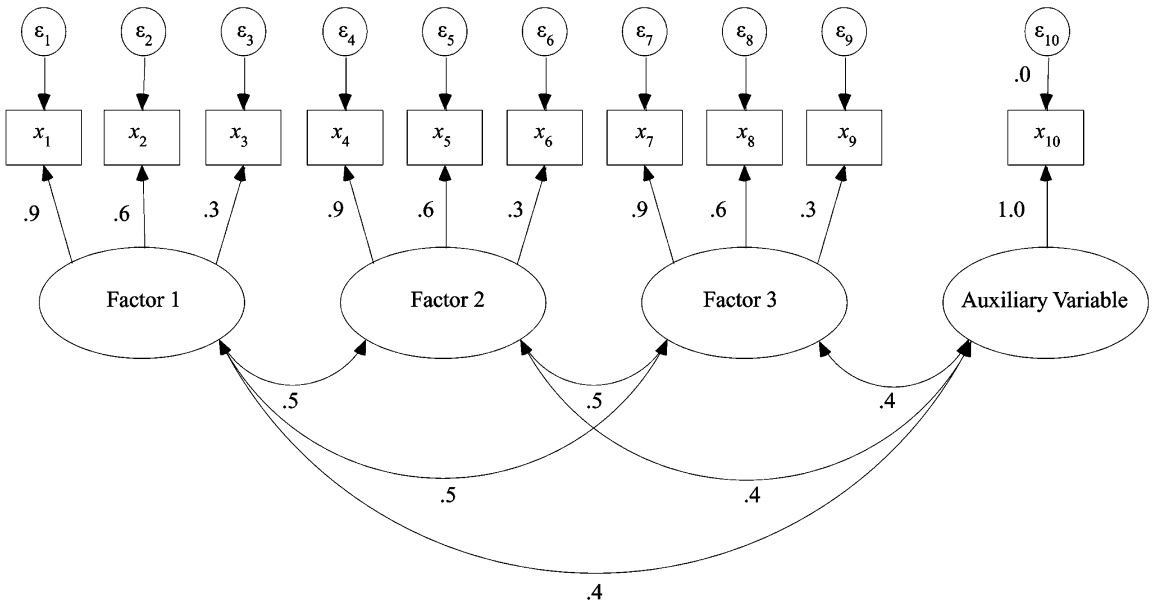

FIGURE 1 Population factor model for the design cells with three factors, three items per factor, between-scale correlations of .50, and high heterogeneity of the within-scale correlations.

Recall that the population model included an additional auxiliary variable that represented the cause of missingness. To impose missing values, each questionnaire item was assigned a corresponding index variable from a uniform distribution with values between zero and one. In the 15% missing data condition, cases in the lower half of the auxiliary variable's distribution were assigned a missing value on a questionnaire item if the corresponding index variable was less than or equal to .30. The same procedure with an index threshold of .10 produced a 5% missing data rate. This produced a situation where low scores on the auxiliary variable were associated with a higher probability of missing data. Because each item had an independent deletion probability, the scale-level missing data rates were different from the item-level rates. Specifically, when the item-level missing data rate was 5%, the scale-level missingness rates were approximately 14% and 46% for the 3 and 12 item scales, respectively. In the 15% missing data condition, the corresponding scale-level missingness rates were roughly 39% and 86%.

Multiple Imputation

We used the PROC MI procedure in SAS 9.1 to implement multiple imputation. For item-level imputation, the imputation model included the individual questionnaire items and the auxiliary variable, whereas the imputation model for scale-level imputation included the scale scores and the auxiliary variable. Consistent with our previous discussion, a scale score was defined as missing when one or more of its items were missing. In line with Graham, Olchowski, and Gilreath (2007), we generated 20 imputed data sets per replication. For each data set, the expectation-maximization (EM) algorithm generated start values for the imputation algorithm, and we set the number of burn-in and between-imputation iterations equal to twice the number of EM iterations or 100, whichever was larger. Finally, we did not force the imputed values to be integers (e.g., by rounding or truncating) and instead used the continuous imputations for all analyses.

Following item-level imputation, we computed a scale score by averaging the items within each scale; this step was not applicable to the scale-level procedure because scale scores were computed prior to imputation. In the analysis phase, we computed a number of scale-level quantities, including means, correlations, and regression coefficients. We used the PROC MIANALYZE procedure in SAS to pool the estimates and standard errors from the 20 data sets into a single set of values (Rubin, 1987). The pooled estimates and symmetric confidence limits provided by MIANALYZE are appropriate for means and regression coefficients because these estimates have a normal sampling distribution. Because the sampling distribution of a correlation coefficient is generally asymmetric, we applied Fisher's r-to-z transformation prior to pooling the estimates and standard

errors. After pooling the correlations and computing confidence intervals on the normalized metric, we back-transformed these quantities to the r metric, thereby producing appropriate asymmetric confidence limits.

Outcomes

We used three criteria to evaluate the performance of each parameter estimate under the various study conditions: bias, mean square error (*MSE*), and confidence interval width. Bias was defined as the difference between the average parameter estimate within a given design cell and the corresponding population parameter. Because categorizing the item responses into 5-point scales made it impossible to analytically derive the population parameters from the factor model parameters, we obtained these values by analyzing a complete-data sample of 500,000 cases within each design cell. *MSE* is the average squared difference between an estimate and the population parameter, as follows:

$$\text{MSE} = \frac{\sum(\hat{\theta} - \theta)^2}{2000}. \tag{3}$$

MSE is an overall measure of accuracy that equals the squared bias of a parameter estimate plus its sampling variance. Finally, confidence interval width was calculated by computing the absolute difference between the upper and lower confidence interval limits. A narrower confidence limit reflects greater precision and statistical power.

Recall that each of the 144 between-subjects design cells had 2,000 replications, resulting in a replication-level sample size of 288,000. With so many replications, nearly any effect would be statistically significant. Consequently, we interpreted only those design effects that exceeded Cohen's (1988) small effect size benchmark (i.e., $\eta^2 = .01$). To maintain comparability of the effect size estimates, we treated imputation strategy (the sole within-subjects factor) as a between-subjects factor when computing η^2 (Dunlap, Cortina, Vaslow, & Burke, 1996). Using 2,000 replications per design cell (or a total replication-level sample size of 576,000) produced very precise effect size estimates. Based on procedures for forming noncentral confidence intervals (Kelley & Maxwell, 2008; Maxwell, Kelley, & Rausch, 2008; Steiger, 2004; Steiger & Fouladi, 1997), the 95% confidence interval width for $\eta^2 = .01$ in our design was approximately .0010.[1] This implies that an η^2 estimate larger than .0105 would produce a

[1]This confidence interval width applies to main effects and interactions with a single degree of freedom and assumes that the remaining design effects collectively explained 6% of the variance (i.e., a medium effect size). Assuming that the remaining effects explained 14% of the variance (i.e., a large effect size) increased the confidence interval width to .0011.

lower confidence interval limit that exceeds the .01 threshold. We believe that this confidence interval width was sufficiently narrow to unambiguously identify effects that exceeded the benchmark of Cohen's (1988) norms for small effect size.

RESULTS

Because the population model in a given design cell had the same factor correlations, the same configuration of factor loadings, and the same missing data rates, there was no reason to expect parameter estimate differences across scales. Consequently, we limit the subsequent presentation to the mean of the first scale, the correlation between the first two scales, and the coefficient from the regression of the first scale on the second scale, controlling for the remaining scales.

Bias

For all scale-level parameter estimates, both item- and scale-level imputation produced trivial biases. As expected, none of the factors or their interactions produced an effect size that exceeded Cohen's (1988) small effect size norms. Consequently, no further discussion of these results is warranted.

Mean Square Error

For the correlation and regression coefficients, the main effect of imputation method was the only design effect that exceeded the small effect size threshold, $\eta^2_{\rho} = .020$ and $\eta^2_{\beta} = .021$. For both parameters, item-level imputation produced the smaller *MSE* and thus greater precision. For the correlation coefficient, the item- and scale-level *MSE* values were 0.0026 and 0.0045, respectively, and the corresponding values for the regression coefficient were 0.0031 and 0.0054, respectively. Because the parameter estimates were unbiased, the *MSE* quantifies only the sampling variance. Consequently, the *MSE* values suggest that scale-level imputation produced slightly less than a twofold increase in the sampling variance of each estimate.

The *MSE* results for the mean were more complex, as a pair of two-way interactions exceeded the small effect size benchmark; the imputation method by number of items interaction, $\eta^2 = .013$; and the imputation approach by missing data rate interaction, $\eta^2 = .011$. In all conditions, item-level imputation produced smaller *MSE* values than scale-level imputation. However, the number of items moderated this effect, such that increasing the items from 3 to 12

decreased the item-level *MSE* and increased the scale-level *MSE*. To illustrate this pattern, Figure 2 shows *MSE* values by imputation method and number of items averaged over the remaining design cells. In addition, the missing data rate was a moderator, such that the disparity between the item- and scale-level *MSE* values increased as the missing data rate increased. Figure 3 illustrates this effect.

As noted previously, the *MSE* values effectively quantify sampling variance. Because this variance is inversely related to sample size, dividing the scale-level *MSE* by the item-level *MSE* gives the proportional increase in the sample size necessary for scale-level imputation to achieve the same precision (i.e., power) as item-level imputation. Recasting the *MSE* values in this fashion demonstrates the practical significance of the previous results. To illustrate, consider the *MSE* ratios for the correlation and the regression coefficient, the values of which were 1.75 and 1.73, respectively (i.e., *MSE* ratio = scale-level *MSE* / item-level *MSE*). All things being equal, these values suggest that a 73 to 75% increase in the sample size would be necessary for scale-level imputation to achieve the same sampling variance (i.e., power) as item-level imputation. The *MSE* ratios for the mean were even more dramatic. Specifically, for the imputation method by number of items interaction, the *MSE* ratios for the 3- and 12-item scales were 1.36 and 3.19. Similarly, for the imputation method by missing data rate interaction, the *MSE* ratios for the 5% and 15% conditions were 1.45 and 2.89.

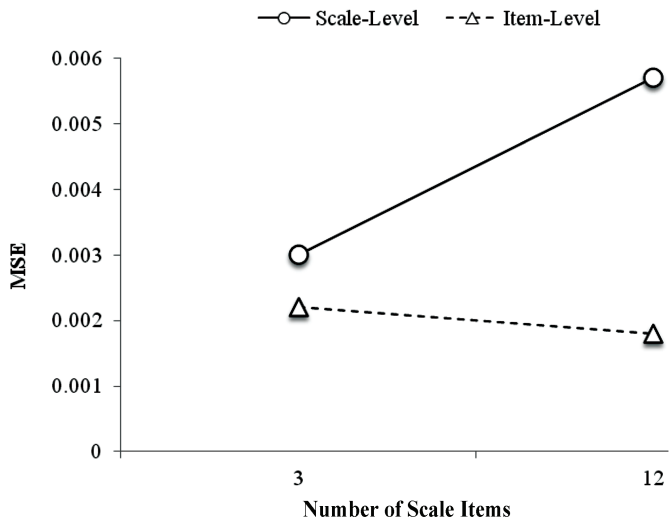

FIGURE 2 Mean squared error (*MSE*) values for the scale mean by imputation method and number of items, averaged over the remaining design cells.

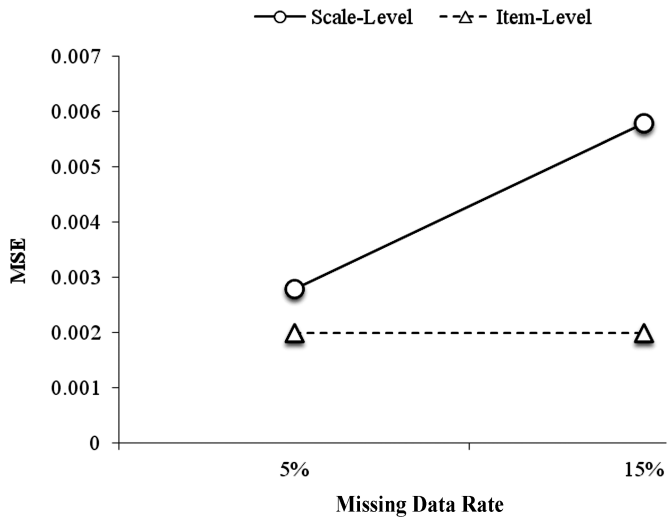

FIGURE 3 Mean squared error (*MSE*) values for the scale mean by imputation method and missing data rate, averaged over the remaining design cells.

Taken as a whole, the *MSE* ratios suggest that researchers can expect item-level imputation to produce a rather dramatic increase in statistical power relative to scale-level imputation.

Confidence Interval Width

The confidence interval results were largely consistent with the previous findings. Considering the confidence interval width of the mean, the two-way interaction between imputation method and sample size exceeded the small effect size benchmark, $\eta^2 = .015$, as did the three-way interaction between imputation method, the number of scale items, and the missing data rate, $\eta^2 = .018$. Across all conditions, item-level imputation produced narrower confidence intervals than scale-level imputation. However, sample size moderated this effect, such that increasing the N had a stronger influence on the width of the scale-level intervals. To illustrate this pattern, Figure 4 shows the mean confidence interval width by imputation method and sample size, averaged over the remaining design cells. For the three-way interaction, increasing the number of scale items increased the disparity between the item- and scale-level interval widths. In turn, the magnitude of this two-way association increased as the missing data rate increased. To illustrate this pattern, the top panel

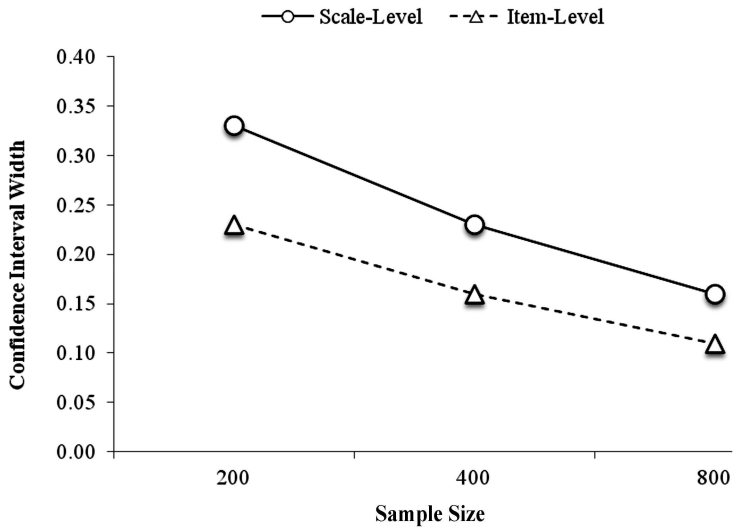

FIGURE 4 Mean confidence interval width for the scale mean by imputation method and sample size, averaged over the remaining design cells.

of Figure 5 shows the mean confidence interval width by imputation method and number of items in the 5% missingness condition, and the bottom panel of the figure shows this two-way association in the 15% missingness condition.

Turning to the correlation and the regression coefficients, the imputation method by sample size interaction exceeded the small effect size threshold, $\eta^2_\rho = .013$ and $\eta^2_\beta = .013$, as did the imputation method by number of items interaction, $\eta^2_\rho = .029$ and $\eta^2_\beta = .025$. The imputation method by missing data rate interaction also explained variation in the confidence interval width of the correlation coefficient, $\eta^2_\rho = .014$. Consistent with the previous results, item-level imputation produced narrower confidence intervals overall. However, sample size moderated this effect, such that increasing the N had a more pronounced impact on the scale-level intervals; this interaction mimicked that in Figure 4. In contrast, increasing the number of scale items increased the disparity between the item- and scale-level interval widths. This effect was similar to that in Figure 5. Finally, increasing the missing data rate also amplified differences between the two imputation methods. Consistent with the *MSE* ratios, the confidence interval results suggest that researchers can expect item-level imputation to produce a rather dramatic increase in statistical power relative to scale-level imputation.

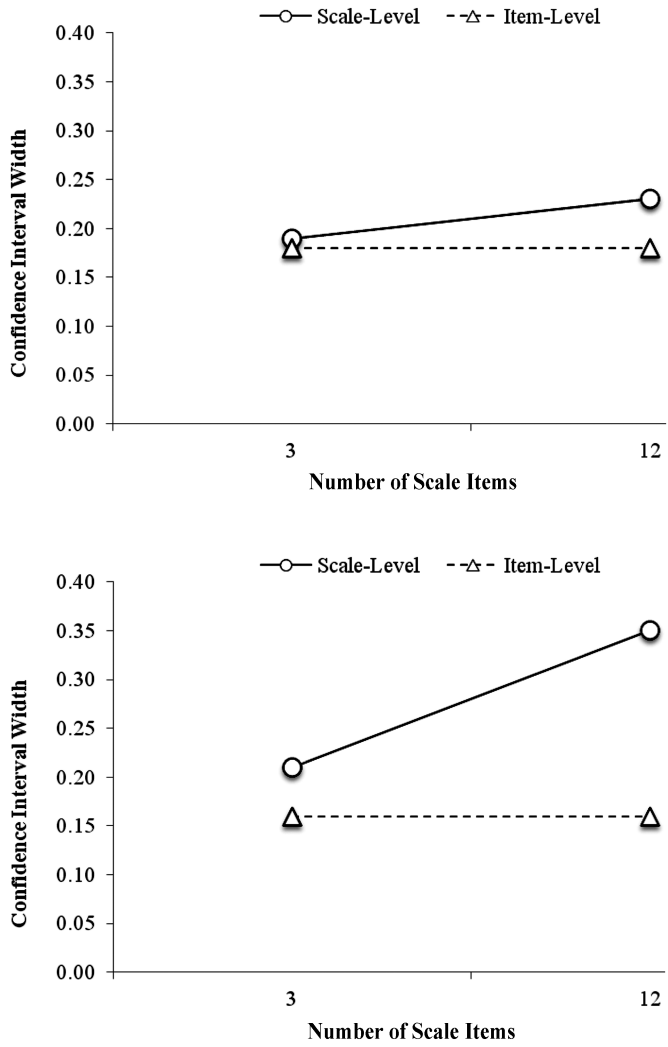

FIGURE 5 Mean confidence interval width for the scale mean by imputation method and number of scale items, averaged over the remaining design cells. The top panel shows the two-way association for the 5% missing data rate condition and the bottom panel shows the corresponding effect for the 15% missing data rate condition.

EMPIRICAL EXAMPLE: ACADEMIC SELF-EFFICACY IN MEXICAN ORIGIN ADOLESCENTS

To illustrate the differences between item- and scale-level imputation, we examined predictors of academic self-efficacy in a sample of 595 Mexican Origin adolescents. The data come from the Bridges to High School Project, a family intervention program designed to prevent school disengagement and mental health disorders for low-income Mexican Origin adolescents (Dillman Carpentier et al., 2007; Gonzales et al., in press). For the purposes of our study, we were not interested in assessing treatment effects and thus limited our analysis to pretest measures that were collected at the start of seventh grade. Our analysis used three multiple-item scales: an effortful control scale that was the mean of sixteen 4-point Likert items (Capaldi & Rothbart, 1992; Ellis & Rothbart, 2001), a general coping efficacy scale that was the mean of eight 4-point Likert items (Sandler, Tein, Mehta, Wolchik, & Ayers, 2000), and an academic self-efficacy scale that was the mean of seven 5-point Likert items (Midgley et al., 2000). The instruments assessing these constructs were validated in previous studies and psychometric analysis confirmed their use in this population (original instruments were shown to be invariant across English and Spanish administration).

Because the Bridges team went to great lengths to avoid missing data, the item-level responses were essentially complete. Rather than creating an artificial missing data mechanism, we imposed missing values according to a planned missing data design where researchers intentionally introduce item-level missingness (Graham et al., 1996; Graham et al., 2006). As seen in Table 1, our design consisted of three test forms or missing data patterns. We randomly assigned participants to forms, such that 40% of the participants were missing 8 of the 16 effortful control items and 2 of the 7 academic self-efficacy items (Test

TABLE 1
Planned Missing Data Design for the Empirical Example

	Item Sets				
	X	*A*	*B*	*C*	
Test Form	EC_1-EC_8 ASE_1-ASE_2	EC_9-EC_{16} ASE_3-ASE_4	GCE_1-CGE_4 ASE_5	GCE_5-CGE_8 ASE_6-ASE_7	*% of Sample*
1	O	M	O	O	40
2	O	O	O	M	40
3	O	O	O	O	20

Note. O = observed; M = missing; EC = effortful control; GCE = general coping efficacy; ASE = academic self-efficacy.

Form 1), 40% of the participants were missing 4 of the 8 general coping efficacy items and 2 of the academic self-efficacy items (Test Form 2), and 20% of the participants had complete data (Test Form 3). Including a pattern with complete data ensured that all bivariate associations were estimable at both the item and the scale level. We implemented this particular design because it allowed us to apply both item- and scale-level imputation (other common designs allow for only type of imputation). Further, the design allowed us to roughly equate the item- and scale-level missing data rates. For example, under an item-level imputation scheme, the 40% of participants who received Test Form 1 had their incomplete effortful control items imputed from the remaining observed items; under scale-level imputation, the same subsample had the incomplete effortful control scale imputed from the general coping efficacy scale. This procedure contrasts our simulation study, where the scale-level missing data rates were larger than the corresponding item-level rates.

We used the SAS MI procedure to generate 20 imputed data sets with 100 burn-in iterations and 100 between-imputation iterations. For item-level imputation, the imputation model included 31 questionnaire items (16 effortful control items, 8 general coping efficacy items, and 7 academic self-efficacy items), whereas the imputation model for scale-level imputation included only the three scale scores (cases with one or more incomplete items had a missing value on the corresponding scale score). In the analysis phase, we estimated the correlations among the scales as well as the regression of the academic self-efficacy scale on the effortful control and general coping efficacy scales. Consistent with the previous computer simulations, we applied Fisher's r-to-z transformation prior to pooling the correlation estimates and standard errors.

Table 2 gives the correlation coefficients and the asymmetric confidence intervals for the three scales. The imputation methods produced similar estimates for two of the three correlations but produced somewhat different estimates of

TABLE 2
Correlation Coefficients from the Empirical Example

Variable	*1.*	*2.*	*3.*
	Item-level imputation		
1. Academic efficacy	1.000	[.275, .451]	[.283, .450]
2. Effortful control	.366	1.000	[.278, .438]
3. Coping efficacy	.369	.361	1.000
	Scale-level imputation		
1. Academic efficacy	1.000	[.269, .581]	[.152, .509]
2. Effortful control	.438	1.000	[.200, .476]
3. Coping efficacy	.343	.345	1.000

Note. The upper diagonal contains asymmetric confidence limits.

the correlation between effort control and academic self-efficacy; presumably, efficiency differences are responsible for this discrepancy. More importantly, item-level imputation produced confidence intervals that were much narrower than those of scale-level imputation. For example, considering the correlation between general coping efficacy and effort control, scale-level imputation produced a confidence interval that was approximately 73% wider than that of item-level imputation (i.e., $[.476 - .200] / [.483 - .278] = 1.725$). Not surprisingly, this translates into a rather dramatic power advantage for item-level imputation.

Turning to the regression of academic self-efficacy on effortful control and general coping efficacy, Table 3 gives the regression coefficients, standard errors, and symmetric confidence intervals for the two imputation approaches; because we centered the predictor variables, the B_0 coefficient estimates the academic self-efficacy mean (Wainer, 2000). Consistent with the computer simulation results, scale-level imputation produced larger standard errors and wider confidence intervals. Surprisingly, the differences between the two approaches exceeded those from the simulation study. For example, consider the effortful control coefficient. Because the squared standard error estimates the sampling variance, the ratio of the squared standard errors is analogous to the *MSE* ratio that we used to evaluate the simulation results. For the effortful control coefficient, this ratio is 2.96, meaning that the sample size would effectively need to be tripled in order for scale-level imputation to produce the same standard error as item-level imputation. Again, the results in Table 3 imply a rather dramatic power advantage for item-level imputation. Perhaps not surprisingly, this power difference also translated into a larger omnibus F test (i.e., the D_1 statistic; Enders, 2010; Schafer, 1997); item-level imputation produced $F(2, 298.85) = 49.37$, $p < .001$, whereas scale-level imputation produced $F(2, 53.198) = 15.82$, $p < .001$.

TABLE 3
Regression Coefficients from the Empirical Example

Parameter	*Est.*	*SE*	*LCL*	*UCL*
	Item-level imputation			
Intercept (B_0)	4.186	0.023	4.140	4.232
Effortful control (B_1)	0.381	0.069	0.244	0.518
Coping efficacy (B_2)	0.328	0.056	0.218	0.439
	Scale-level imputation			
Intercept (B_0)	4.233	0.047	4.135	4.332
Effortful control (B_1)	0.451	0.119	0.203	0.699
Coping efficacy (B_2)	0.233	0.100	0.026	0.440

Note. LCL and UCL = lower and upper 95% confidence intervals.

DISCUSSION

To date, methodologists have devoted little attention to the differences between item- and scale-level imputation for incomplete questionnaire data. This is an important practical issue given the frequency with which behavioral science researchers use scale score analyses to address their substantive questions. Although item-level imputation is intuitively superior because it incorporates stronger correlates of the incomplete variables, the extent to which a richer imputation model impacts scale-level parameter estimates is unclear. Evaluating the relative performance of the two imputation approaches is also important because there are situations where item-level imputation is difficult or impossible. For example, in data sets with many questionnaire items (e.g., a longitudinal study where researchers administer multiple questionnaires at each wave), the item-level imputation model may be prohibitively large, leading to convergence problems. In these situations, scale-level imputation may be a sensible compromise (Graham, 2009). Design-based considerations can also necessitate scale-level imputation. Such is the case when, for logistical reasons, a researcher implements a planned missing data design that keeps scale items together in the same questionnaire booklet. Given the lack of existing research, we used Monte Carlo simulations to assess the impact of imputation method on the bias and efficiency (i.e., power) of scale-level parameter estimates (e.g., scale score means, between-scale correlations, and regression coefficients).

As expected, the imputation approach had no material impact on bias. We designed the simulations to mimic a situation where an auxiliary variable was responsible for missingness. Statistical theory predicts that multiple imputation should yield consistent parameter estimates when the auxiliary variable is part of the imputation model, which it was. However, it is worth noting that other deletion mechanisms could produce different conclusions. For example, when the propensity for missing data on a particular scale item is related to another item from the same scale, scale-level imputation could introduce bias, whereas item-level imputation should not. Whether such a scenario would introduce appreciable bias to scale-level parameter estimates is an open question for future research.

Our primary goal was to assess the relative efficiency (i.e., power) of item- and scale-level imputation. As discussed previously, item- and scale-level imputation differ in the number and the quality of the predictor variables that they incorporate into the imputation process (see Equations 1 and 2). Although we expected item-level imputation to produce smaller standard errors and narrower confidence intervals, it was impossible to predict the magnitude of this effect for a scale-level analysis. The *MSE* ratios from the simulation were surprisingly large, suggesting that item-level imputation had a dramatic impact on scale-level statistical power. For measures of association, scale-level imputation required a 75% increase in

the sample size to achieve the same sampling variance (i.e., squared standard error) as item-level imputation. The differences from the empirical sample were even more extreme. Perhaps not surprisingly, the benefit of item-level imputation increased as the number of questionnaire items increased. Considered as a whole, the simulation results and the empirical example suggest that researchers should adopt item-level imputation whenever possible. Of course, this recommendation is limited to situations where the sample size is large enough to support item-level imputation.

Our study also has implications for the application of planned missing data designs. For logistical reasons, a researcher might prefer a design that keeps a set of scale items together on the same test form (Graham et al., 2006, p. 327). Returning to Table 1, such a design would result if the X set consisted of demographic and background variables, the A set consisted of the academic self-efficacy questionnaire, and the B and C sets consisted of the general coping efficacy and effortful control scales, respectively. It is important to note that this design precludes the possibility of item-level imputation because a set of scale items is either completely missing or completely observed. To date, the methodological literature provides relatively little guidance on the distribution of individual questionnaire items across test forms. In fact, the advice has changed somewhat over the years; Graham et al. (1996) recommended a split-scale approach that distributes questionnaire items across item sets (e.g., the design in Table 1), whereas Graham et al. (2006) recommended keeping items together on the same test form. Both the computer simulation results and the empirical illustration suggest that the latter approach can have a detrimental impact on power given that it necessitates scale-level imputation. The existing methodological literature tends to emphasize the logistical details of planned missingness designs rather than the analysis of data from such designs. Our results suggest that researchers need to carefully consider analytic issues when designing planned missingness studies.

Limitations

Although we attempted to maximize generalization of the present results by choosing experimental conditions that were representative of published research papers, our simulation nevertheless has limitations that are important to consider. First, computational and storage considerations forced us to limit the number of items in our population model (the raw data files from the simulation required more than one terabyte of drive space, and it took multiple computers several months to generate, impute, and analyze the raw data). This is an important issue because item-level imputation becomes difficult or impossible when the number of items gets very large. To illustrate, consider a longitudinal study where a researcher administers six 12-item questionnaires on five occasions.

Not including demographic variables or other auxiliary information, item-level imputation would require a sample size capable of supporting a 360-variable regression model; at a minimum, the sample size must equal or exceed the number of variables, but incomplete data sets may support far fewer variables. Establishing useful guidelines for model complexity is difficult because a variety of data-specific features influence the convergence behavior of iterative imputation algorithms (e.g., the correlations among the variables, missing data patterns, missing data rates, and item distribution shapes; Schafer, 1997). Nevertheless, future studies should investigate the behavior of item-level imputation as participant-to-variable ratios approach unity.

In situations where the number of items exceeds the sample size, implementing a ridge prior distribution can often alleviate convergence difficulties (Enders, 2010, pp. 256–259; Schafer, 1997, pp. 155–157). Conceptually, the ridge prior increases the effective sample size by adding imaginary data points from a population where the items are uncorrelated but have the same mean and variance as the sample data. This strategy may allow researchers to employ item-level imputation as the participant-to-variable ratio approaches one, but it does so at the cost of attenuating measures of association. However, this bias is often negligible, particularly when the prior distribution contributes a small number of additional degrees of freedom (i.e., imaginary data points) to estimation. Although we did not investigate boundary conditions or the ridge prior in our study, these would be interesting possibilities for future research. In situations where the number of items greatly exceeds the sample size, future studies should focus on alternative strategies that allow researchers to retain item-level information during the imputation process. We outline a few possibilities in a subsequent section.

The missing data pattern that we implemented in our simulation is a second limitation worth considering. In our program, an auxiliary variable determined missingness, such that each questionnaire item had an independent deletion probability (e.g., a participant refuses to respond to certain items but not others). This is in contrast to a situation where entire blocks of questionnaire items are concurrently missing (e.g., a participant skips an entire page of items, a researcher implements a planned missing data design such as that in Table 1). This design choice meant that the number of within-scale predictors in the item-level imputation model varied across cases (e.g., for some cases, a missing item was imputed from a single item; for other cases a missing item was imputed from several other items) and the scale-level missing data rate exceeded the item-level rate (i.e., because the probability of missing one or more item responses increased with the number of items). Because the missing data pattern impacts the relative performance of the two imputation approaches, our simulation results may not generalize to other scenarios with item-level missingness. In an attempt to address this concern, we used the empirical example to explore

the performance of the imputation approaches in a planned missingness design with a very different missing data pattern and approximately equal item- and scale-level missing data rates. Although the conclusions were the same (item-level imputation was superior), the efficiency differences from the empirical data generally exceeded those from the simulation study. Future studies should investigate the missing data pattern's influence on the relative performance of these two approaches.

Potential Solutions and Future Directions

Based on our results, the magnitude of the efficiency gain from item-level imputation casts serious doubt on the utility of scale-level imputation. Nevertheless, the sample size requirement for item-level imputation is a very real concern for many researchers—recall that the number of cases must exceed the number of variables, perhaps by a substantial amount. In this section, we outline potential solutions for dealing with item-level missing data in large imputation problems. In particular, we outline two sequential imputation approaches that split a large imputation problem into a series of smaller imputation problems. It is important to point out that these approaches are currently ad hoc solutions, although we believe that they can be formalized as Markov Chain Monte Carlo algorithms within the Bayesian framework.

As a springboard for discussion, consider a longitudinal study where a researcher administers six 12-item questionnaires on five occasions (i.e., 360 questionnaire items, not including demographic variables or other auxiliary information). This is a prototypical situation where researchers might have difficulty implementing item-level imputation. At first glance, it might seem sensible to perform item-level imputation separately at each wave. However, doing so makes the implicit and unrealistic assumption that the repeated measures variables are uncorrelated across time because the imputation model for a particular wave omits information from other assessments. A solution is to perform item-level imputation at each wave while using the scale scores from previous assessments as auxiliary variables. The procedural steps are as follows: (a) create m sets of item-level imputations at Wave 1 and compute the scale scores for each imputed data set; (b) append a copy of the Wave 2 data to each of the m sets of scale scores; and (c) for each of the m appended data sets, generate a single set of item-level imputations at the next wave. Repeatedly applying Steps (b) and (c) generates a full set of item-level imputations, and including the scale scores from earlier assessments (e.g., Wave 3 imputation would incorporate the scale scores from Waves 1 and 2) preserves between-wave associations.

The previous strategy is useful, for example, if a researcher wanted to begin analyzing the data from earlier waves prior to collecting data at later waves. Little, McConnell, Howard, and Stump (2008) outlined a comparable procedure

that uses scale scores from all questionnaires (or in the longitudinal example, all waves) to preserve the data structure. This so-called three-step approach for item-level imputation is as follows: (a) create a complete set of placeholder scale scores (e.g., using scale-level imputation or by averaging the available items within each scale), (b) sequentially perform item-level imputation on each questionnaire while using the placeholder scale scores from all other questionnaires as auxiliary variables, and (c) delete the placeholder scale scores and compute new scale scores from the item-level imputations. Note that the Little et al. (2008) procedure is applicable to longitudinal or cross-sectional designs. As a final option, researchers could employ one of several single imputation techniques designed for item-level missing data. Perhaps the most popular method in the behavioral sciences is to compute a scale score by averaging the available item responses for each individual. For example, if a respondent answered 8 out of 10 questionnaire items, her scale score would be the mean of the 8 observed items. The literature refers to this as a prorated scale score or person mean imputation (the method is algebraically equivalent to imputing the items with each individual's mean response). Despite its popularity, prorating has no theoretical justification; nor does it have a body of empirical literature supporting its use (see Downey & King, 1998; Enders, 2003). In practice, computing a prorated scale score probably works best when the item means are similar and the between-item correlations are uniform in magnitude (Graham, 2009; Schafer & Graham, 2002). Unfortunately, it is difficult to assess these characteristics prior to imputation. As an alternative to prorating, methodologists have developed imputation techniques that combine within- and between-person variation to generate filled-in item responses (Bernaards & Sijtsma, 1999, 2000; Van Ginkel et al., 2007a, 2007b). Van Ginkel (2010) provides an overview and an evaluation of these techniques. Although these methods lack a strong theoretical rationale, they have thus far performed well in simulation studies.

As a caveat, all single imputation techniques attenuate standard errors, even if they yield accurate parameter estimates. Analyzing a single data set effectively treats the imputations as real data and provides no mechanism for estimating the additional sampling error that results from missing data (analyzing multiply imputed data sets provides such an adjustment). Significance tests from these procedures should be viewed with caution unless a researcher is willing to employ additional computational procedures to estimate standard errors (e.g., the bootstrap; Enders, 2010, pp. 145–148).

CONCLUSION

A researcher applying multiple imputation to incomplete questionnaire data can either impute the incomplete items prior to computing scale scores or

impute the scale scores directly from other scale scores. Although the choice of imputation approach may not impact the accuracy of scale-level parameter estimates, it certainly can impact the efficiency of the estimates. Item-level imputation consistently produced a meaningful power advantage, the magnitude of which casts serious doubt on the utility of scale-level imputation. Because item-level imputation can require relatively large samples, it is important to develop formal imputation algorithms that can accommodate large numbers of questionnaire items.

ACKNOWLEDGMENTS

Amanda C. Gottschall is now at the Parenting & Family Research Center, University of South Carolina. We thank Nancy A. Gonzales for allowing us to use the Bridges data in our empirical example.

REFERENCES

Beck, A. T., Ward, C., & Mendelson, M. (1961). Beck Depression Inventory (BDI). *Archives of General Psychiatry, 4,* 561–571.

Belin, T. R., Datt, M., Desmond, K., & Ganz, P. A. (1999). Comparing imputation of entire subscales versus individual items in a study of quality of life following breast cancer. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 813–818.

Bernaards, C. A., & Sijtsma, K. (1999). Factor analysis of multidimensional polytomous item response data suffering from ignorable item nonresponse. *Multivariate Behavioral Research, 34,* 277–313.

Bernaards, C. A., & Sijtsma, K. (2000). Influence of imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. *Multivariate Behavioral Research, 35,* 321–364.

Capaldi, D. M., & Rothbart, M. K. (1992). Development and validation of an early adolescent temperament measure. *Journal of Early Adolescence, 12,* 153–173.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

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

Dillman Carpentier, F. R., Mauricio, A. M., Gonzales, N. A., Millsap, R. E., Meza, C. M., Dumka, L. E., . . . Genalo, M. T. (2007). Engaging Mexican Origin families in a school-based preventive intervention. *Journal of Primary Prevention, 28,* 521–546.

Downey, R. G., & King, C. V. (1998). Missing data in Likert ratings: A comparison of replacement methods. *Journal of General Psychology, 125,* 175–191.

Dunlap, W. P., Cortina, J. M., Vaslow, J. G., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 1*, 170–177.

Ellis, L. K., & Rothbart, M. K. (2001, April). *Revision of the early adolescent temperament questionnaire.* Poster session presented at the biennial meeting of Society for Research in Child Development, Minneapolis, MN.

Enders, C. K. (2003). Using the expectation maximization algorithm to estimate coefficient alpha for scales with item-level missing data. *Psychological Methods, 8,* 322–337.

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

Gmel, G. (2001). Imputation of missing values in the case of a multiple item instrument measuring alcohol consumption. *Statistics in Medicine, 20,* 2369–2381.

Gonzales, N. A., Dumka, L. E., Millsap, R. E., Gottschall, A., McClain, D. B., Wong, J. J., … Kim, S. Y. (in press). Randomized trial of a broad preventive intervention for Mexican American adolescents. *Journal of Consulting and Clinical Psychology.*

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60,* 549–576.

Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research, 31,* 197–218.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8,* 206–213.

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

Kaiser, H. E., & Dickman, K. (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika, 27,* 179–182.

Kelley, K., & Maxwell, S. E. (2008). Sample size planning with applications to multiple regression: Power and accuracy for omnibus and targeted effects. In P. Alasuutari, L. Bickman, & J. Brannen (Eds.), *The Sage handbook of social research methods* (pp. 166–192). London, UK: Sage.

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

Little, T. D., McConnell, E. K., Howard, W. J., & Stump, K. N. (2008). *Missing data in large data projects: Two methods of missing data imputation when working with large data projects* (KUant Guide No. 011.3). Retrieved from University of Kansas, KUant guides website: http://www.quant.ku.edu/pdf/11._Imputation_with_Large_Data_Sets.pdf

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59,* 537–563.

McDonald, R. A., Thurston, P. W., & Nelson, M. R. (2000). A Monte Carlo study of missing item methods. *Organizational Research Methods, 3,* 71–92.

Midgley, C., Maehr, M. L., Hruda, L. Z., Anderman, E., Anderman, L., Freeman, K. E., … Urdan, T. (2000). *Manual for the patterns of adaptive learning scale.* Ann Arbor: University of Michigan.

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

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

Sandler, I. N., Tein, J.-Y., Mehta, P., Wolchik, S., & Ayers, T. (2000). Coping efficacy and psychological problems of children of divorce. *Child Development, 71,* 1099–1118.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York, NY: Chapman & Hall/CRC.

Schafer, J. L. (2003). Multiple imputation in multivariate problems where the imputer's and analyst's models differ. *Statistica Neerlandica, 57,* 19–35.

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.

Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods, 6,* 317–329.

Steiger, J. H. (2004). Beyond the *F* test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, *9*, 164–182.

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical methods. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 221–257). Mahwah, NJ: Erlbaum.

Van Ginkel, J. R. (2010). Investigation of multiple imputation in low-quality questionnaire data. *Multivariate Behavioral Research, 45,* 574–598.

Van Ginkel, J. R., Van der Ark, L. A., & Sijtsma, K. (2007a). Multiple imputation for item scores when test data are factorially complex. *British Journal of Mathematical and Statistical Psychology, 60,* 315–337.

Van Ginkel, J. R., Van der Ark, L. A., & Sijtsma, K. (2007b). Multiple imputation of test and questionnaire data and influence on psychometric results. *Multivariate Behavioral Research, 42,* 387–414.

Wainer, H. (2000). The centercept: An estimable and meaningful regression parameter. *Psychological Science, 11,* 434–436.