MULTIVARIATE BEHAVIORAL RESEARCH
2016, VOL. 51, NOS. 2–3, 296–313
http://dx.doi.org/10.1080/00273171.2016.1144502

Routledge
Taylor & Francis Group

Accounting for Parcel-Allocation Variability in Practice: Combining Sources of Uncertainty and Choosing the Number of Allocations

Sonya K. Sterba and Jason D. Rights

Vanderbilt University

ABSTRACT

Item parceling remains widely used under conditions that can lead to parcel-allocation variability in results. Hence, researchers may be interested in quantifying and accounting for parcel-allocation variability within sample. To do so in practice, three key issues need to be addressed. First, how can we combine sources of uncertainty arising from sampling variability and parcel-allocation variability when drawing inferences about parameters in structural equation models? Second, on what basis can we choose the number of repeated item-to-parcel allocations within sample? Third, how can we diagnose and report proportions of total variability per estimate arising due to parcel-allocation variability versus sampling variability? This article addresses these three methodological issues. Developments are illustrated using simulated and empirical examples, and software for implementing them is provided.

KEYWORDS

Item parceling; parcel-allocation variability; pooling rules; structural equation modeling; sampling variability

Psychologists commonly create parcel scores by summing or averaging subsets of items and then use the parcel scores as factor indicators in structural equation models (SEMs). Reviews suggest that between one out of five (Bandalos & Finney, 2001; Hall, Snell, & Foust, 1999) and one out of two (Plummer, 2000; Williams & O'Boyle, 2008) SEM applications use parceling. Parceling is often employed and recommended in the common context of low item communalities and/or small samples (e.g., Bagozzi & Edwards, 1998; Matsunaga, 2008; Meade & Kroustalis, 2006; Nasser-Abu & Wisenbaker, 2006; West, Finch, & Curran, 1995; Williams & O'Boyle, 2008; Yang, Nay, & Hoyle, 2010; Yuan, Bentler, & Kano, 1997). In this context, parceling also has been used to avoid potential estimation problems that could be encountered when employing categorical variable estimation methods with many ordered-categorical items (see Bandalos, 2008; West et al., 1995; Yang et al., 2010). Reasons for these recommendations include the fact that, compared with item-level models, parcel-level models have higher communalities and fewer estimated parameters, and may have lower risk of convergence problems (e.g., Little, Cunningham, Shahar, & Widaman, 2002; Little, Rhemtulla, Gibson, & Schoemann, 2013).

Alternative item-to-parcel allocations within sample (i.e., alternative parceling strategies) were long considered to provide the same results in terms of structural parameters and model fit so long as items loading on a given factor were unidimensional in the population (e.g., Hagtvet & Nasser, 2004; Hall et al., 1999; Hau & Marsh, 2004; Landis, Beal, & Tesluk, 2000; Little et al., 2002; Marsh, Lüdtke, Nagengast, Morin, & Van Davier, 2013; Matsunaga, 2008; Meade & Kroustalis, 2006; Nasser-Abu & Wisenbaker, 2006; Plummer, 2000; Rogers & Schmitt, 2004; Sass & Smith, 2006; Williams & O'Boyle, 2008). However, recent research has shown that in this context, model fit, standard errors, and parameter estimates—including those for structural parameters—can vary meaningfully across alternative allocations of items to parcels, given a fixed number of items per parcel and parcels per factor (Sterba, 2011; Sterba & MacCallum, 2010). This result holds even in the context of equal item loadings on each factor. The amount of such *parcel-allocation variability* in results is elevated under the exact conditions where parceling has been widely recommended (e.g., small samples and/or modest communalities). In addition, if items loading on a given factor are not unidimensional in the population, further parcel-allocation variability would arise (e.g., Bandalos, 2002; Bandalos & Finney, 2001). Since SEM applications using parceling typically assume unidimensional items for each factor in the population (see Marsh et al., 2013), in this article we conservatively focus on the context where parcel-allocation variability arises despite item unidimensionality in the population.

CONTACT Sonya K. Sterba Sonya.Sterba@vanderbilt.edu Department of Psychology and Human Development, Vanderbilt University, Peabody #552, 230 Appleton Place, Nashville, TN 37203, USA
Color versions of one or more of the figures can be found online at www.tandfonline.com/hmbr.

In this context, consider a common scenario where a researcher is interested in fitting a two-factor confirmatory factor analysis (CFA) model with 15 items per factor, using the combination scheme of five parcels per factor and three items per parcel. There are 2.82×10^{16} possible item-to-parcel allocations that could be chosen within sample under this combination scheme. However, researchers typically pick just one allocation of items to parcels (i.e., one parceling strategy). Often this allocation is chosen randomly; other times it is substantively justified. Researchers typically do not consider what results would have been obtained from other potential allocations. Even if a researcher has a substantive justification for employing a particular allocation, the researcher may still question how generalizable the results are with respect to results that could be obtained from other allocations within the sample. For example, Cattell and Burdsal (1975) expressed concern that a particular choice of parcel allocation could be "too subjective" and could depend on "stereotypes of a particular experimenter" (p. 167). Under some data/model conditions, alternative parcel allocations give very similar results. Under other data/model conditions, alternative parcel allocations can give very different results (e.g., parameter estimates ranging from significant to nonsignificant). In summary, researchers may wonder whether their results are representative of those that could have been obtained from the distribution of possible parcel allocations within sample.

To address such concerns about generalizability and representativeness of results across alternative parcel-allocations, researchers need to be able to quantify the variability in results across alternative parcel allocations within sample. Sterba and MacCallum (2010) suggested a reporting strategy to accomplish this. Specifically, they suggested that a researcher interested in random item-to-parcel allocating could repeatedly randomly allocate items to parcels 100 times, within his or her single sample. Then the researcher could fit his or her SEM model to each allocation and report the following information about the within-sample parcel-allocation distribution of each parameter estimate and standard error:

- across-allocations mean, standard deviation, minimum, and maximum of each parameter estimate,
- across-allocations mean, standard deviation, minimum, and maximum of each standard error, and
- proportion of allocations in which the null hypothesis could be rejected for each parameter.

This reporting strategy was intended to communicate uncertainty due to parcel-allocation variability in results for each parameter. However, there are three important limitations to this reporting strategy: (1) Parameter estimate and standard error results were not pooled across allocations to yield a single inferential decision per parameter; (2) the number of repeated allocations was chosen arbitrarily; and (3) the amount of variability per parameter estimate due to sampling variability versus parcel-allocation variability was not distinguished and quantified. The purpose of the current article is to address these three limitations. Each limitation is described further, in conjunction with how it will be addressed in the current article.

How can we pool results across allocations within sample, per each parameter?

The first limitation of Sterba and MacCallum's (2010) reporting strategy is that they did not provide a single test per parameter that incorporated uncertainty due to *both* sampling variability and parcel-allocation variability. Therefore, their strategy did not facilitate a single decision regarding whether to reject the null hypothesis for that parameter. For example, in the empirical example supplied to illustrate this strategy, the effect of an *agreeableness* latent factor on a *tangible support* latent factor was significant in 78% of allocations within sample and nonsignificant in 22% of allocations within sample. What overall substantive conclusion should the researcher draw from this result? Is 78% of allocations so many that the researcher ought to reject the null hypothesis? What if the null hypothesis had been rejected in 10% of allocations within sample? Or 50% of allocations within sample? The fundamental issue is that the previously recommended reporting strategy did not provide a way to *combine* estimates across allocations and *combine* standard errors across allocations to make a single inferential decision about rejecting the null hypothesis per parameter. Furthermore, the previously recommended reporting strategy was not compact; it required nine columns of results to summarize parcel-allocation variability in a set of parameter estimates and standard errors. The *first goal* of the current article is to provide and demonstrate an approach, from Rubin (1987), for pooling parameter estimates and pooling standard errors across allocations to yield a single inferential decision per parameter that reflects two sources of variability: across-samples and across-allocations within-sample. Not only will this approach provide a single null hypothesis significance test per parameter (rather than 100 of them), but we will also extend previous research by providing a pooled confidence interval for each parameter.

How should we choose the number of repeated allocations?

The second limitation of Sterba and MacCallum's (2010) reporting strategy had to do with the chosen number of

repeated item-to-parcel allocations within sample. The number of allocations chosen (100) was arbitrary. Of course, it would not be practical computationally to repeat the SEM analysis for every single allocation from the finite population of millions of allocations possible using a given combination scheme. Instead, researchers may want to gauge how many allocations are required to obtain pooled parameter estimate and standard error results that are reasonably stable—even in the event that a specified number of new allocations from that combination scheme are analyzed. There is reason to anticipate that the number of allocations needed to fulfill this objective would differ depending on particular model and data conditions, rather than being a fixed number (e.g., 100). The *second goal* of the current article is to address the limitation of an arbitrarily selected number of repeated random allocations. Specifically, we develop and illustrate a computational algorithm for choosing the number of allocations, in a given sample, that are needed to obtain a desired degree of stability in pooled parameter estimates and standard errors that are reflective of both sampling and parcel-allocation variability.

How can we diagnose the amount of parameter-estimate variability due to sampling versus parcel allocating?

A third limitation of the previous reporting strategy is that it provided no method to quantify the increase in such pooled standard errors due to the presence of parcel-allocation variability, above and beyond the presence of sampling variability. This increase would not necessarily be the same for each parameter in a given model. Some parameters' standard errors may be little affected by parcel-allocation variability, whereas other parameters' standard errors may be greatly affected. It could be substantively useful to know which parameters' standard errors are more affected. The *third goal* of the current article is to provide indices that (a) quantify the contribution of parcel-allocation variability to the total variability in each estimate and (b) compare the amount of parcel-allocation variability versus sampling variability in each estimate.

Broadly speaking, the goals of this article reflect a common statistical interest in accounting for variability arising from both sampling and nonsampling sources when interpreting results (e.g., Bayarri et al., 2007; Cole, Chu, & Greenland, 2006; Ghosh-Dastidar & Schafer, 2003; Groves & Lyberg, 2010; Hoeting, Madigan, Raftery, & Volinsky, 1999; MacCallum, 2013; Mislevy, 1991; Mislevy, Johnson, & Muraki, 1992; Reiter & Raghunathan, 2007; Rubin, 1987; Stuart & Rubin, 2008). The focus of this article on pooling estimates across allocations within sample also reflects the *principle of aggregation* (see discussion in Little et al., 2013; Matsunaga, 2008; Nunnally, 1978; Rushton, Brainerd, & Pressley, 1983). That is, results from any single parcel allocation are less representative of the entire within-sample parcel-allocation distribution than are results produced by combining information across allocations.

The remainder of this article proceeds as follows. First, we motivate and describe the use of Rubin's (1987) rules as an approach for pooling estimates and standard errors for SEM parameters across random allocations within sample. We show that this approach matches results of a Monte Carlo pooling approach. We also describe hypothesis testing and confidence intervals for SEM parameters. Second, we describe an algorithm to select the number of allocations needed to produce pooled estimates and standard errors that maintain a desired degree of stability when a specified number of new allocations are included. We also motivate and then investigate hypotheses regarding data and model features expected to, on average, increase or decrease the number of allocations required to meet these convergence criteria. Third, we describe indices to assess the proportion of variance in a parameter estimate that is due to allocation variability. Fourth, we demonstrate the application of methods introduced in prior sections in an empirical example. This example makes use of software tools developed in R to allow researchers to obtain results described in previous sections. We conclude with a discussion of future research topics. In this regard, note that the scope of the present article pertains to parameter estimates and standard errors, rather than to model fit; the latter is discussed as a future direction.

Pooling parameter estimates and pooling standard errors across allocations within sample

Pooling using Rubin's rules

Rubin (1987) proposed rules for combining two sources of variability—arising from repeated-sampling and nonsampling sources—when drawing inferences about a given parameter. In Rubin's (1987) context, the nonsampling variability arose due to missing data. Specifically, there was uncertainty about what the missing values of variables would have been, had they been observed. Multiple plausible scores (i.e., imputations) for these missing values were drawn from a distribution of missing values given observed values, and then for each ($m = 1 \ldots M$) draw, a complete-data analysis was performed. Finally, the M sets of parameter estimates and standard errors were combined using these pooling rules (for more details see, e.g., Enders, 2010; Little & Rubin, 2002).

Subsequently, Rubin's (1987) pooling rules have been used extensively outside the missing data context to combine other nonsampling sources of variability, together with sampling variability, when drawing inferences (see Reiter & Raghunathan 2007, for review). For example, these pooling rules have been used to combine sampling variability with uncertainty in latent ability scores in educational testing (e.g., Li, 2012; Mislevy, 1991; Mislevy et al., 1992), sampling variability with uncertainty in matches from multiple control groups or multiple data sets (e.g., Rässler, 2003; Stuart & Rubin, 2008), or sampling variability with uncertainty due to measurement/response error generally (Asparouhov & Muthén, 2010a; Cole et al., 2006; Ghosh-Dastidar & Schafer, 2003).

In the present context, the sources of variability to be combined in creating pooled estimates and standard errors arise from both repeated sampling and repeated parcel allocating within sample. Using this approach, a pooled parameter of inferential interest is defined as the expected value of the parcel-level parameter, across repeated allocations within sample and across repeated samples. For pooled structural parameters, this is simply the structural parameter from the generating item-level model (Sterba & MacCallum, 2010). In many empirical parceling applications, substantive interest lies only in structural parameters (Bandalos & Finney, 2001; Little et al., 2002, 2013; Marsh et al., 2013; Matsunaga, 2008; Meade & Kroustalis, 2006; Plummer, 2000; Rogers & Schmitt, 2004; Sass & Smith, 2006; Sterba & Rights, in press; Stucky, Goffredson, & Panter, 2012; Williams & O'Boyle, 2008). If substantive interest also lies in parcel-level measurement parameters, and unidimensionality and random parcel-allocating are assumed, each pooled parcel-loading parameter is the average generating item loading across all items per factor (see the Appendix for details), and each pooled parcel residual variance parameter is the average generating item residual variance divided by the number of items per parcel (see the Appendix for details). Of course, if researchers are interested in making inferences about individual item-level loadings and item-level residual variances, they are advised to instead fit an item-level SEM rather than a parcel-level SEM (Bandalos & Finney, 2001; Little et al., 2002; Matsunaga, 2008; Meade & Kroustalis, 2006; Plummer, 2000; Rogers & Schmitt, 2004; Sass & Smith, 2006; Stucky et al., 2012; Williams & O'Boyle, 2008). Furthermore, if researchers are not comfortable assuming unidimensional items on a given factor and are interested in exploratory analyses to compare competing measurement models, they likewise are advised to fit an item-level SEM, rather than a parcel-level SEM (see Bandalos, 2002, 2008; Bandalos & Finney, 2001; Hall et al., 1999; Hagtvet & Nasser, 2004; Little et al., 2002; Matsunaga, 2008; Meade & Kroustalis, 2006; Plummer, 2000; Rogers & Schmitt, 2004; Sass & Smith, 2006; Stucky et al., 2012; Williams & O'Boyle, 2008).

Rubin (1987, 1996) justified the pooling rules using both randomization-based (frequentist) and Bayesian arguments. Here, we relate Rubin's (1987) randomization-based frequentist arguments for use of the pooling rules to the present context of pooling estimates and standard errors across random allocations within sample, under the assumption that items loading on each factor are unidimensional. We begin with some definitions.

A *combination scheme* for repeated item-to-parcel allocations within sample is uniquely defined by the number of factors, K, with parcel indicators (where factors are indexed $k = 1 \ldots K$); the number of item indicators for the kth factor, r_k; the number of parcels for the kth factor, p_k; and the number of items per parcel j of factor k, q_{jk}. We refer to a generic combination scheme as $C_{q_{jk}p_k}$. For such a combination scheme, the total possible number of item-to-parcel allocations within a given sample is

$$T = \prod_{k=1}^{K} \left(\frac{r_k!}{\prod_{j=1}^{p_k} (q_{jk}!)} \right). \quad (1)$$

From a randomization-based perspective, the total number of allocations possible using a particular combination scheme represents an allocation *frame* (e.g., Kish, 1965; see Sterba, 2009, for review). Each draw or allocation, denoted A, from this frame has a known, nonzero probability of selection. Specifically, in our case of repeated random parcel allocating, each allocation has a $(1/T)$ probability of selection (i.e., *simple random* allocating). Let M indicate the size of a set of allocations randomly[1] selected from the frame within a researcher's single sample (where allocations are indexed $m = 1 \ldots M$). Later in this article, we address choosing the number M. Here, repeated allocations are drawn *with replacement*, implying that a given allocation can in theory be selected more than once in a set.

Suppose a SEM is fit to each allocation $m = 1 \ldots M$ within sample. Most commonly, this would be done using maximum likelihood estimation. Rubin's (1987) rules assume that estimates from the mth fitted model

[1] Although we are conceptualizing allocations as randomly selected, by *repeatedly* randomly allocating we are likely to encounter, by chance, allocations that fit the description of certain existing purposive parceling strategies that can assume unidimensionality (see Little et al., 2013, or Matsunaga, 2008, for review). For example, one purposive strategy that could be encountered by chance is a correlational strategy (Rogers & Schmitt, 2004). Using this strategy for a given factor, parcels are initially seeded with pairs of the most highly correlated items, and subsequent items are assigned to the parcel with which they are most highly correlated.

are normally distributed across repeated samples. Consistent with this assumption, for the mth allocation's fitted model, maximum likelihood estimates are asymptotically normally distributed across repeated samples (for review, see Bollen, 1989). Under simple random allocating with replacement, Equation (2) provides an estimate of the expected value of the model parameter θ across samples and allocations generated using combination scheme $C_{q_{jk}p_k}$.

$$\bar{\theta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m. \tag{2}$$

For large M, $\bar{\theta}$ will be approximately normally distributed due to the central limit theorem (Rubin, 1987, 1996). Under simple random allocating with-replacement, Equation (3) is an estimate of the variance of the $\hat{\theta}$ across samples and across allocations generated using combination scheme $C_{q_{jk}p_k}$.

$$V_T = V_W + V_B + \frac{V_B}{M}. \tag{3}$$

The term V_W in Equation (3) quantifies uncertainty due to sampling variability. The term V_W is defined in Equation (4) as the across-allocation average of the repeated-sampling variance of $\hat{\theta}$. In allocation m, the repeated sampling variance of $\hat{\theta}$ is the square of its analytic standard error, denoted SE_m^2.

$$V_W = \frac{1}{M}\sum_{m=1}^{M}SE_m^2. \tag{4}$$

The term V_B in Equation (3) quantifies uncertainty due to parcel-allocation variability. The term V_B is defined in Equation (5) as the between-allocation variance of $\hat{\theta}$.

$$V_B = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\theta}_m - \bar{\theta})^2. \tag{5}$$

The final term in Equation (3), V_B/M, is an estimate of the variance of $\bar{\theta}$ across repeated sets of M allocations (Schafer, 1997), and the influence of this term disappears as M increases. The pooled standard error is thus

$$SE_{pool} = \sqrt{V_T}. \tag{6}$$

More information on the frequentist/randomization-based justification for Equations (2)–(5) is given in Rubin, (1987, 1996), Schafer (1997), and Van Buuren (2012).

Demonstration of the correspondence between Rubin's rules and a Monte Carlo approach for pooling estimates and SE

To make the interpretation of Equations (2)–(5) and the assumptions underlying them concrete in the present context, it is instructive to compare the convenient analytic pooling approach using Rubin's (1987) rules to an alternative Monte Carlo pooling approach that entails the same assumptions. First, we describe this Monte Carlo pooling approach, and then we provide a demonstration that the within-sample results obtained using Rubin's pooling rules match those obtained using the alternative Monte Carlo approach.

Like Rubin's (1987) rules, for allocation m, the Monte Carlo approach assumes normality for the estimate's repeated sampling distribution. Like Rubin's (1987) rules, the Monte Carlo approach does not make any distributional assumptions about the parcel-allocation distribution (across m) other than to require simple random allocating with replacement. The Monte Carlo pooling approach involves implementing the following four steps:

Step 1. For each allocation $m = 1 \ldots M$, fit the parcel-level SEM. Record the vector of parameter estimates from allocation m. Also record the asymptotic covariance matrix of the parameter estimates from allocation m.

Step 2. For each allocation $m = 1 \ldots M$, generate a sampling distribution of multivariate normal estimates using the mth allocation's estimate vector as the generating mean vector and the mth allocation's asymptotic covariance matrix as the generating covariance matrix. For example, the sampling distribution in allocation m might consist of $R = 1{,}000$ generated estimates.

Step 3. For each model parameter, pool all $M \times R$ (e.g., 100×1000) estimates generated in Step 2. This creates a distribution pooled across repeated samples and repeated parcel allocations, which is conditional on the items from the original sample.

Step 4. For each model parameter, compute the mean and standard deviation of the pooled sampling/allocation distribution from Step 3. The mean of the pooled distribution should closely match Rubin's pooled estimate in Equation (2). The standard deviation of this pooled distribution should closely match Rubin's pooled standard error in Equation (6).

Next, we demonstrate the correspondence of within-sample results obtained using Rubin's pooling rules and using the Monte Carlo pooling approach. This demonstration uses a simulated data set of $N = 100$ generated

Table 1. Parameter estimates and standard errors pooled across allocations within a single sample using Rubin's (1987) rules versus a Monte Carlo approach.

	Pooling using Rubin's (1987) rules		Pooling using Monte Carlo approach	
	Avg Est	Pooled *SE*	Avg Est	Pooled *SE*
λ_{11}	.480	.090	.480	.090
λ_{12}	.496	.082	.496	.082
λ_{13}	.499	.088	.499	.087
λ_{14}	.490	.082	.490	.082
λ_{15}	.498	.086	.498	.085
λ_{21}	.548	.110	.548	.110
λ_{22}	.535	.106	.535	.105
λ_{23}	.558	.101	.558	.101
λ_{24}	.546	.110	.546	.109
λ_{25}	.543	.106	.543	.105
θ_{11}	.237	.060	.237	.060
θ_{12}	.227	.058	.227	.058
θ_{13}	.221	.060	.221	.060
θ_{14}	.227	.056	.227	.056
θ_{15}	.228	.060	.228	.060
θ_{21}	.241	.069	.241	.069
θ_{22}	.248	.065	.248	.065
θ_{23}	.235	.070	.235	.069
θ_{24}	.240	.071	.240	.071
θ_{25}	.250	.066	.250	.066
ϕ_{12}	.279	.110	.279	.110

Note. *SE* = Standard error. Est = estimate. Avg = average.

from a CFA model with $K = 2$ factors and 15 unidimensional item indicators of each factor in the population.[2] Within this sample, we randomly allocate items to parcels $M = 100$ times for each factor using a combination scheme where there are $p_k = 5$ parcels for each factor and $q_{jk} = 3$ items per parcel j of factor k. Then, we fit our parcel-level two-factor CFA model to each of the $M = 100$ allocation-specific data sets, within sample, using maximum likelihood estimation.

Note that for identification, we fixed each factor variance to 1 (and gave loadings positive start values). However, we could have instead fixed one loading to 1 on each factor; if so, the other loadings and the factor variance would be rescaled accordingly.[3]

We record the parameter estimates and standard errors from the M allocations. The pooled estimates and pooled standard errors computed using Rubin's (1987) rules (Equations [2]–[6]) are given in the left-hand two columns of Table 1. These results closely match those computed using the Monte Carlo approach, in the right-hand columns of Table 1.

[2] In generating the item-level model, the item factor loadings alternated among $\lambda = .4, .5, .6$, and corresponding item residual variances alternated among $\theta = .84, .75, .64$ so that items had unit variances. Factor variances and covariances were $\phi_{11} = 1$, $\phi_{22} = 1$, $\phi_{12} = .25$. Parameters that are the subject of inference in a parcel-level analysis were discussed earlier in the article text and do not include individual item-level loadings and residual variances.

[3] Structural parameters are estimated under slightly different assumptions across factor identification methods; they will not necessarily yield the same pooled *z*-statistic results across factor identification methods for reasons discussed in, for example, Gonzalez and Griffin (2001).

Note that some parameters' sampling distributions will not be normal in small samples (e.g., correlations or variances) even if they are asymptotically normally distributed. Despite this fact, in the missing data context, multiple imputation software in widespread use routinely implements Rubin's (1987) rules in the pooling phase for all model parameters, under normality assumptions for parameter estimates' sampling distributions (for review, see Enders, 2010; Harel & Zhou, 2007). Applied researchers fitting SEMs thus widely use pooled results under this assumption in practice (Asparouhov & Muthén, 2010b). It is possible to apply a normalizing transformation to a parameter estimate prior to the application of Rubin's rules or prior to the application of the alternative Monte Carlo approach (e.g., Ratitch, Lipkovich, & O'Kelly, 2013; Van Buuren, 2012). Hypothesis testing could then be performed using the standard error for the transformed estimate. Such a preliminary transformation would not affect the correspondence between the Rubin's rules analytic approach and the Monte Carlo approach. For instance, in our example we could transform ϕ_{12} (which was shown to have a pooled Est = .279 and pooled SE = .110 using either pooling approach, in Table 1) to a z' metric using Fisher's (1921) transformation. Fisher's transformation can be implemented as a model constraint in our M fitted CFA models. Doing so, we still obtain the same estimate and pooled standard error using both Rubin's rules and the Monte Carlo pooling approach, now in the z' metric (pooled Est = .287, pooled SE = .119, with either approach). When computing confidence interval bounds for such a parameter, as described in the next section, the upper and lower bounds could be computed in the z' metric, and then both bounds could be converted back to the original parameter's (e.g., correlation) metric.

In summary, in this subsection, the Monte Carlo approach was introduced to illustrate the assumptions under which Rubin's (1987) rules are applied in the parcel-allocation context. The within-allocation distributional assumptions under which Rubin's rules are applied in the parceling context are the same as the within-imputation distributional assumptions under which Rubin's rules are currently widely applied in the missing-data context. We anticipate that, in practice, researchers would typically be interested in using the simpler Rubin's rules approach, rather than the Monte Carlo approach, to compute estimates and standard errors pooling sampling and parcel-allocation variability. Thus, we have implemented Rubin's analytic approach in an R program for use with repeated random allocations from a given combination scheme, as described later.

Null hypothesis testing and confidence intervals for pooled parameters

Now that we have discussed computing pooled parameter estimates and pooled standard errors across allocations, it is of interest to discuss inference for each parameter. In SEM applications, null hypothesis significance tests (NHSTs) for individual parameters are typically z tests (e.g., Bollen, 1989). Accordingly, in the context of missing data, when NHSTs are conducted for individual SEM parameters using multiple imputation, often z tests are employed (e.g., Muthén & Muthén, 1998–2016). In the context of parcel allocations, we could similarly employ z tests[4] for each model parameter, with the test statistic calculated using quantities from Equations (2) and (6): $\hat{\bar{\theta}}/\widehat{SE}_{pool} \sim N(0, 1)$. For example, these tests could be of interpretive interest for testing the null hypothesis that a pooled factor correlation is 0 or that a pooled parcel-loading is 0 in the population.

Confidence intervals (*CI*s) for each pooled parameter can also be calculated as $CI_{100(1-\alpha)} = \hat{\bar{\theta}} \pm z_{\alpha/2}\widehat{SE}_{pool}$. A 95% *CI*, for example, encloses 95% of the pooled-sampling and repeated-allocations distribution. If our null hypothesized value, θ_{null}, lies outside this CI_{95}, then we can reject the null hypothesis. We can thus conclude that, across repeated samples and repeated allocations within sample, we are unlikely to observe such a large pooled estimate if the null hypothesis is true. Thus, our inference regarding pooled parameter estimates incorporates uncertainty due to repeated sampling and repeated parcel-allocating. To illustrate, we reanalyzed the empirical example from Sterba and MacCallum (2010) using their original 100 allocations. For the slope mentioned earlier (effect of the *tangible support* factor on the *agreeableness* factor), pooled results now provide a single inferential decision: $\hat{\bar{\theta}} = -.45,\ \widehat{SE}_{pool} = .24,\ p = .06,$ and $CI_{95}\{-.91, .02\}$.

Note that a wide *CI* corresponds with less precision and corresponds with greater variance of the pooled distribution. However, simply knowing that the *CI* is wide does not inform us whether more of the variability in the estimate is due to sampling variability or due to parcel-allocation variability. In a later section, we consider how to assess the relative contributions of parcel-allocation variability versus sampling variability to SE_{pool}.

[4] Rubin (1987) also considers t tests $\hat{\bar{\theta}}/\widehat{SE}_{pool} \sim t(df)$ where $df = (M-1)(1+(M\hat{V}_W)/(M\hat{V}_B+\hat{V}_B))^2$. We do not discuss these here since t tests are less standard when fitting SEMs and since this *df* increases with *M*; at the large *M* needed in the context of parcel allocating (discussed later; see, e.g., Table 2) the corresponding *p* value closely matches that of a *z* test. Nonetheless, NHST and CI using both *t* and *z* reference distributions are provided for each model parameter in our R function.

Choosing the number of item-to-parcel allocations within sample

The previous section described methods for pooling parameter estimates and standard errors across M allocations within sample to yield a single inferential decision per parameter. As mentioned previously, we are interested in randomly drawing a set of M allocations rather than taking a census of all the, for example, millions or trillions of possible allocations under a particular combination scheme. In this section, we address how to choose this number of allocations, M.

To motivate the task of choosing M, we consider the behavior of the pooled parameter estimates and pooled SE as M increases, within sample. Figure 1 plots variability in pooled parameter estimates (Panel A) and pooled standard errors (Panel B) across 50 repeated *sets* of randomly drawn allocations, each of size $M = 5$, 10, 25, 50, 100, 150, 200, 300, or 400 for a single sample. This $N = 100$ sample was generated from the two-factor CFA described in Footnote 2, and each line in Figure 1 corresponds to a different parcel-level parameter. Figure 1 shows that, as the size of M increases, there is a reduction of variability in results, should a different set of allocations be randomly chosen. Note that the rate of descent in the Figure 1 plots will be particular to this model and these data conditions. We are thus motivated to pick an M that leads to less variability in results, and we may need to do so in a manner that is tailored to specific model and data conditions.

Relatedly, existing literature has addressed choosing the number of imputations in the missing data context and choosing the number of plausible values in other data analysis contexts. Recent studies have indicated that the optimal number of imputations or plausible values depends to some extent on the statistic or quantity of interest (point estimate, standard error, etc.), the objective (e.g., stability of results despite further increases in the number of plausible values), and the model/data conditions (e.g., sample size, percent of missing data) (Asparouhov & Muthén, 2010a; Bodner, 2008; Graham, Olchowski, & Gilreath, 2007; Reiter, 2007; White, Royston, & Wood, 2011). For some objectives, three to five imputations may be sufficient (see Rubin, 1987). For other objectives, many imputations may be required. For example, recent simulations have shown that ≥ 100 imputations may be necessary if the objective is for pooled standard errors to remain stable despite further increases in the number of imputations (e.g., Graham et al., 2007; also see Bodner, 2008; White et al., 2011). A large number of imputations is necessary for the latter objective because the between-imputation variance (Equation [5]), which appears in the pooled standard error computation, stabilizes as the number of imputations increases. Authors of

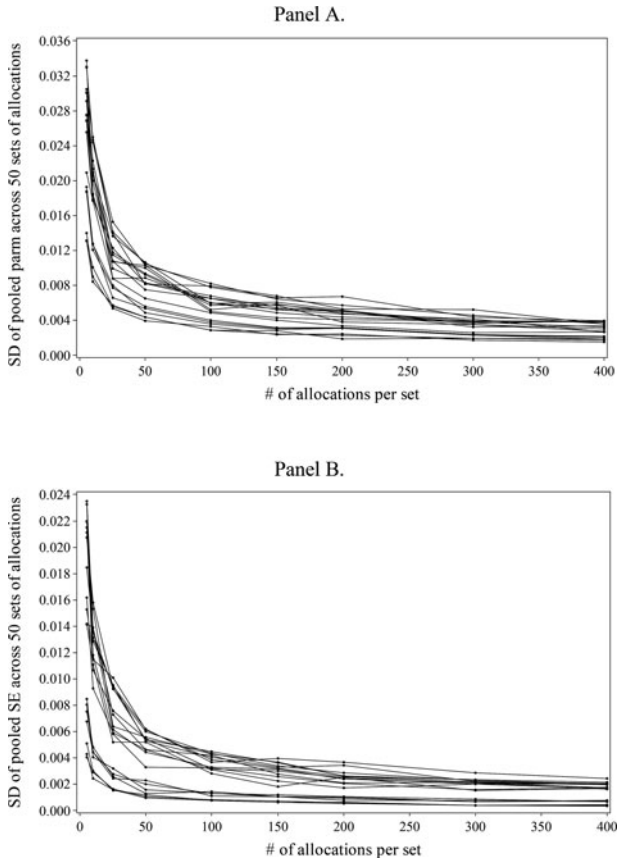

Figure 1. Variability of pooled parameter estimates (Panel A) and pooled standard errors (Panel B) across 50 *sets* of *M* allocations within a single sample, where the size of *M* (on the *x*-axis) ranges from $M = 5$ to $M = 400$. See text for a description of the generating and fitted models. For a given choice of *M* (*x*-axis value), a set of *M* random item-to-parcel allocations is drawn and used to create *M* parcel-level data sets. These data sets are separately analyzed and results are pooled. This process is repeated for 50 different *sets* of *M* random allocations, at each size of *M*. The *y*-axis value depicts the standard deviation of results across these 50 sets of *M* random allocations. Each line corresponds to pooled results for a particular parameter. There are as many lines per plot as there are freely estimated parameters. Specifically, in Panel A, each line depicts how the variability in the pooled estimate of a particular parameter, across repeated *sets* of *M* allocations, decreases as the size of *M* increases. In Panel B, each line depicts how the variability in the pooled standard error of a particular parameter, across repeated *sets* of *M* allocations, decreases as the size of *M* increases. SD = standard deviation. SE = standard error.

these simulations also cautioned that the results regarding sufficient numbers of imputations could be specific to their chosen model/data conditions (e.g., a very simple one-parameter model was used in Bodner (2008) that would not mirror models used in most applications).

Similarly, in the parcel-allocation context, different objectives could be defined, leading to different optimal numbers of allocations. Here, we define the following objective in choosing the number of allocations *M*. In the discussion, we describe other potential objectives that could be of future research interest.

> Objective: To select a sufficiently large *M* such that, if a specified greater number of allocations were randomly drawn ($M* = M + M_{\text{inc}}$), pooled parameter estimates and standard errors would not change appreciably.

This stated objective clarifies that our interest is in finding *M* that yields pooled parameter estimates and pooled standard errors whose values show robustness to a specified increase in the number of allocations. Note that, in this stated objective, the phrase "would not change appreciably" refers to our convergence criteria (defined subsequently). We can anticipate that data/model conditions that would lead to more parcel-allocation variability would also lead to larger *M* needed to fulfill this objective. As discussed later, these conditions that lead to more parcel-allocation variability should include smaller sample size, fewer items per parcel and parcels per factor, and more factors (holding constant the number of items per parcel and parcels per factor; Sterba, 2011; Sterba & MacCallum, 2010). The relationship between these conditions and necessary *M* will be investigated subsequently. Presently, we focus on pragmatics of how to choose *M* to fulfill our stated objective.

In the missing data context, the number of imputations is sometimes chosen using rules of thumb, based on interpolation/extrapolation from existing simulation results, but is other times chosen using algorithms tailored to a particular set of data and model conditions (Bodner, 2008; Graham et al., 2007; Reiter, 2007; Royston, 2004; White et al., 2011). Here we consider both strategies, starting with the latter. To choose *M* that fulfills the above objective, we developed and implemented an iterative algorithm that can be adapted and used with a researcher's item-level data set and SEM. In the Discussion, we consider when this algorithm would be most useful versus when a rule-of-thumb for the number of allocations could be sufficient.

Iterative algorithm for choosing the number of allocations (M)

To employ the algorithm requires first specifying a combination scheme (Equation [1]) for generating random allocations; specifying a starting/baseline number of allocations, M_{start}; and specifying a number of allocations by which to increment, M_{inc}. In this article, we conservatively use $M_{start} = 5$ and $M_{inc} = 5$ because preliminary investigations indicated that large values (e.g., 50) could lead to more variability in the final choice of *M* within sample. The algorithm implements the following steps:

1. At iteration $h = 1$, the SEM is fit to M_{start} randomly drawn allocations, and pooled parameter estimates and pooled standard errors are computed using Equations (2)–(6).
2. At iteration $h > 1$, the SEM is fit to $M_{start} + ((h - 1) \times M_{inc})$ brand new randomly drawn allocations. Pooled parameter estimates and pooled standard errors are computed using Equations (2)–(6). This step is repeated until convergence criteria are met.

We define the algorithm's *convergence criteria* as follows. Convergence is met when each pooled parameter estimate and each pooled standard error changes by less than δ_a% of its value at the previous iteration (or changes less than δ_b, *if* δ_a% is smaller than δ_b). When implementing the algorithm in practice, researchers have the option to specify their own choice of δ_a and δ_b, based on their own substantive considerations. However, we use $\delta_a = 1\%$ and $\delta_b = .01$ in this article because smaller values are unlikely to be substantively meaningful in most contexts.[5] These convergence criteria are examples of the maximum absolute deviation type of convergence criteria (Thisted, 1988). If the convergence criteria are met at the iteration of the algorithm using $M + M_{inc}$ allocations, this implies that the algorithm converges to the choice of M allocations.

Researchers have the option of interpreting pooled parameter estimates and pooled standard errors using those M allocations, as will be done here. Researchers also have the option of outputting results using all available allocations employed by the algorithm in determining M (i.e., $\sum_{h=1}^{H} (M_{start} + (h - 1) \times M_{inc})$ allocations), where H is the total number of iterations until convergence. Under this option, allocations from the previous iteration are not discarded at the beginning of Step 2. We do not demonstrate this option here, however, because our focus is on illustrating the behavior and performance of the algorithm at the chosen M.

Our convergence criteria concern pooled estimates and pooled standard errors (rather than other statistics) because our objective in choosing M specifically concerns achieving a desired degree of stability in these quantities. To fulfill this objective, our convergence criteria need to include pooled standard errors, despite the fact that the final term in Equation (3) itself involves M. This final term in Equation (3) should become inconsequential as M increases; furthermore, it is relevant to consider when this happens (similarly to Bodner, 2008; Graham et al., 2007; White et al., 2011).

[5] Note also that it is possible to just use δ_a (i.e., set $\delta_b = 0$). However, we feel it is useful to specify $\delta_b = .01$ to accommodate very small pooled values for which the $\delta_a = 1\%$ criterion would be extremely small.

It is possible to specify that only a *subset* of freely estimated parameters contributes toward determining convergence. That is, the convergence criteria needn't be applied to the entire set of all pooled parameter estimates and pooled standard errors if there is a substantive reason to exclude some of them from determining convergence. For example, if certain parameter estimates will not be substantively interpreted in a given application (e.g., intercepts), they could potentially be excluded from the convergence criteria and thus not contribute toward determining the choice of M.

Demonstration of the algorithm for choosing M within sample

To illustrate the application of this algorithm within sample, here we use it to choose M in a simulated sample of $N = 100$ generated from a $K = 2$ factor item-level CFA model. In the population, there were 15 normally distributed, unidimensional item indicators of each factor. Item factor loadings were all $\lambda = .4$, and item residual variances were all $\theta = .84$ so that items had unit variances. The fact that loadings are equal is not relevant to the illustration and was done for simplicity. Factor variances and covariances were $\phi_{11} = 1$, $\phi_{22} = 1$, $\phi_{12} = .25$. To create each parcel-level data set, an item-to-parcel allocation was drawn from a combination scheme where there were $p_k = 5$ parcels per factor with $q_{jk} = 3$ items per parcel. This implies a communality of .36 for each parcel indicator in the fitted parcel-level models. A two-factor parcel-level CFA model was fit to each parcel-level data set. The algorithm was specified to have $M_{start} = 5$ and $M_{inc} = 5$.

The top three panels of Figure 2 plot pooled estimates (for different kinds of parameters) versus iteration (i.e., loop) of the algorithm, within sample. The bottom three panels of Figure 2 plot the associated pooled standard errors (for each kind of parameter) versus iteration of the algorithm, within sample. Researchers mainly interested in interpreting structural parameters in the parcel-solution would focus on the middle column of plots, where the single jagged line corresponds to the pooled estimate (top) and pooled standard error (bottom) results for the factor correlation. The first column of plots contains five jagged lines per plot; in this first column, each line corresponds to results for a particular loading. For clarity of presentation, results for the five loadings from only one of the two factors are superimposed in a given plot. Results for residual variances from the same factor are plotted in the third column; in this third column, each line corresponds to results for a particular residual variance. Results from the other factor exhibited the same pattern. The grey vertical line indicates the iteration (x-axis value) at which the global convergence criteria

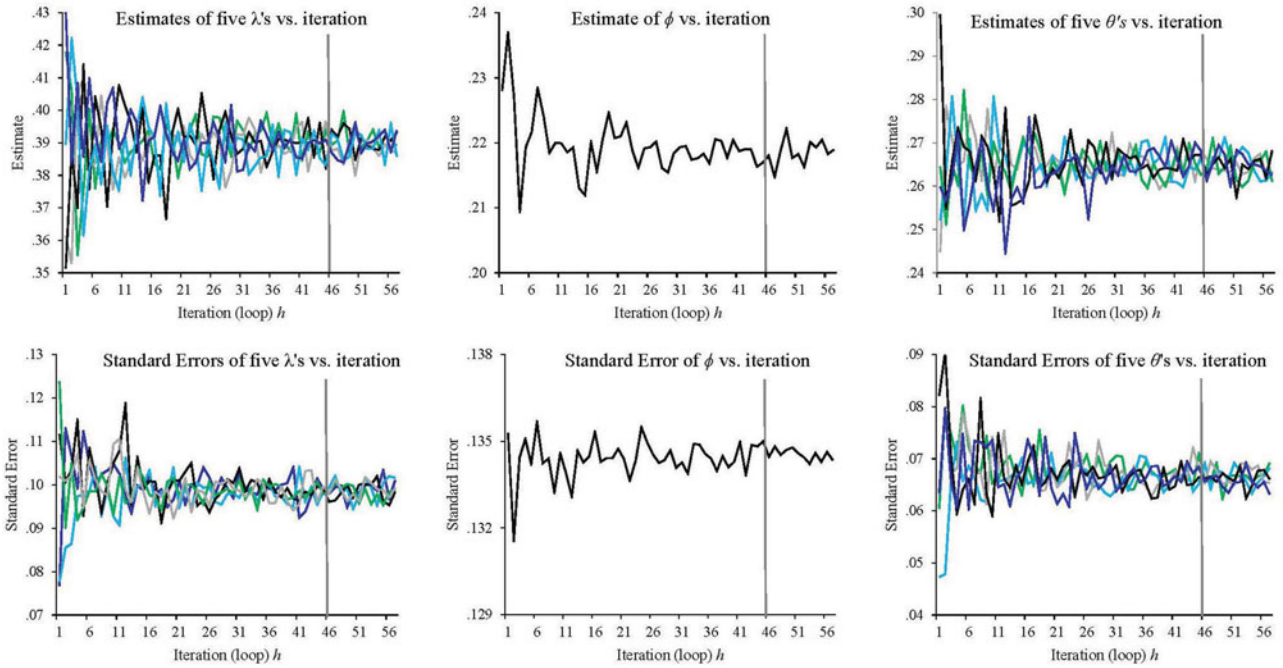

Figure 2. Plots of pooled parameter estimates versus iteration of the algorithm for choosing *M* and plots of pooled standard errors versus iteration of the algorithm for choosing *M*: Illustrative CFA application in a single sample. Iteration 1 (i.e., loop 1) uses five allocations drawn with replacement from the allocation scheme (i.e., $M_{start} = 5$). At each subsequent iteration, the number of allocations increments by five (i.e., $M_{inc} = 5$), and all allocations are redrawn with replacement. The grey vertical line is where the algorithm's global convergence criteria are met (here, at $M = 235$ allocations). Only 5 of 10 loadings (λ) and 5 of 10 residual variances (θ) from the two-factor CFA model are shown for parsimony. ϕ = factor covariance. Each jagged line corresponds with a single parameter's results. M = the number of random item-to-parcel allocations that meets the user-specified convergence criteria (see text) for the algorithm. CFA = confirmatory factor analysis model. h = iteration (i.e., loop) of the algorithm (see text).

were met for all pooled estimates and pooled standard errors simultaneously (iteration $H = 46$, where $M = 235$ allocations). Pooled parameter estimates and standard errors could be interpreted at this *M*. Figure 2 shows that, for low numbers of iterations (i.e., when there are few allocations total), pooled parameter estimates and pooled standard errors each bounce around considerably when allocations are redrawn with-replacement at the subsequent iteration. This uncertainty diminishes as the number of iterations increases (and thus *M* increases). Pooled parameter estimates and standard errors can still exhibit limited variation after the convergence criteria were met.

The algorithm's convergence criteria require local stability of results (i.e., between consecutive iterations). In Figure 2, results for 10 iterations post-convergence are shown to the right of the grey vertical lines. Future research could use diagnostics to assess the stationarity of results after the convergence criteria are met (for example, an adaptation of Gelman & Rubin, 1992).

Data/model conditions leading to choice of larger M

In the previous section, a demonstration of the algorithm for choosing *M* was conducted using one model and one data set under one set of conditions. More generally, certain data/model conditions may tend to necessitate smaller or larger *M* to fulfill the algorithm's objective (defined earlier) of meeting local stability criteria for pooled parameter estimates and pooled standard errors. The central limit theorem implies that, as *M* increases, the stability in the between-allocation variance (Equation [5]) increases. Relatedly, data/model conditions leading to larger parcel-allocation variability (larger between-allocation variance) may tend to require larger *M* to reach stability. Conditions that increase parcel-allocation variability (Sterba, 2011; Sterba & MacCallum, 2010) and thus may increase required *M* include: smaller sample size (i.e., more sampling error) as well as the combination of fewer items per parcel and fewer parcels per factor (i.e., less information per indicator combined with less well overdetermined factors). For a given number of items per parcel and parcels per factor, more factors should also lead to larger required *M* because of the greater number of estimated parameters that need to simultaneously reach stability to fulfill the convergence criteria, particularly if measurement parameters are included in the convergence criteria. To demonstrate the relationship between required *M* and these conditions, a simulation study was conducted. One hundred samples were generated from

Table 2. Comparison of selected *M*, on average across 100 samples, for different model and data conditions.

		Effect of sample size	
Sample size	100	250	
#factors	2	2	
#items/parcel, #parcels/factor	3, 5	3, 5	
M: Average (*SD*)	145 (34)	70 (18)	
	Effect of #factors†		
Sample size	100	100	
#factors	2	4	
#items/parcel, #parcels/factor	3, 5	3, 5	
M: Average (*SD*)	145 (34)	217 (45)	
		Effect of #items/parcel, #parcels/factor	
Sample size	100	100	100
#factors	2	2	2
#items/parcel, #parcels/factor*	3, 5	5, 3	3, 3
M: Average (*SD*)	145 (34)	78 (25)	187 (89)

Note. M: Average = number of allocations chosen within sample by the iterative algorithm, averaged across 100 samples. *M: SD* = standard deviation of chosen *M* across 100 samples.

∗ Note that the 3,5 and 5,3 conditions imply 15 items/factor, and the 3,3 condition implies 9 items/factor.

† Note that the number of factors was increased, holding constant the number of items/parcel and number of parcels/factor.

CFA models under each of several conditions. Conditions were as follows:

$N = 100$, factors $= 2$, items/parcel $= 3$, parcels/factor $= 5$
$N = 250$, factors $= 2$, items/parcel $= 3$, parcels/factor $= 5$
$N = 100$, factors $= 4$, items/parcel $= 3$, parcels/factor $= 5$
$N = 100$, factors $= 2$, items/parcel $= 3$, parcels/factor $= 3$
$N = 100$, factors $= 2$, items/parcel $= 5$, parcels/factor $= 3$

Generating λ and θ parameters in this simulation matched the simulated example from the previous section. Factor variances were all 1. Factor covariances were $\phi_{12} = .25$ in the two-factor condition, and $\phi_{12} = .25$; $\phi_{13} = .50$; $\phi_{14} = .10$; $\phi_{23} = .20$; $\phi_{24} = .40$; $\phi_{34} = .15$ in the four-factor condition.

In each condition, the number of allocations, *M*, was selected within sample using the algorithm; all parameters except for indicator intercepts were used in determining convergence. The average chosen *M* across the 100 repeated samples per condition was computed. Table 2 compares the average chosen *M*, across conditions. Table 2 indicates that, on average, *M* is twice as large for $N = 100$ than for $N = 250$, all else equal. In addition, *M* is on average 50% larger for four factors than for two factors. Furthermore, holding constant the number of item indicators of a factor (15), fewer larger parcels led to lower *M* than did more smaller parcels (average $M =$ 78 vs. 145). The combination of fewer items per parcel and fewer parcels per factor led to greater average *M* (i.e., 187) across samples.

In summary, certain data/model conditions can, on average, systematically increase or decrease the necessary number of allocations *M*. For this reason, it can be useful for researchers to use an algorithm to select *M* under their specific data/model conditions, which may differ from those studied here. An R program implementing the proposed algorithm is available on the authors' websites and incorporated into a release of the *semTools* R package (see Footnote 7).

Indices to quantify uncertainty in estimates due to sampling versus allocation variability

Suppose a researcher has chosen *M* using procedures in the previous section. Also suppose the researcher has computed and interpreted pooled parameter estimates and pooled standard errors at that *M* (as described in the first section). In addition, the researcher may want an index of the proportion of total variability in each parameter estimate that is due to parcel-allocation variability. Furthermore, the researcher may also want an index of the amount of parcel-allocation variability relative to sampling variability in each parameter estimate. These two indices are discussed in the current section. For each parameter estimate, these two indices can be computed from the total variance in Equation (3), within-allocation variance in Equation (4), and between-allocation variance in Equation (5).

The first index we consider is the proportion of the total variance of a parameter estimate that is attributable to parcel-allocation variability (PPAV). The PPAV can be calculated for each model parameter estimate as follows:

$$PPAV = \frac{V_B + (V_B/M)}{V_T}. \quad (7)$$

Note that this formula is also used in the multiple imputation literature involving missing data (e.g., Enders, 2010; Savalei & Rhemtulla, 2012; Schafer & Graham, 2002; also see closely related expressions in Schafer, 1997; Schafer & Olson, 1998) to assess the contribution of between-imputation variability to uncertainty about the parameter. The PPAV in Equation (7) can be interpreted as an effect size, ranging from 0 to 1, that measures the contribution of parcel-allocation variability to the total variance of the estimate (akin to an R^2; Enders, 2010). If there is no parcel-allocation variability in a particular parameter estimate, Equation (7) will be 0. If Equation (7) is .30, then 30% of the total variability in that estimate is contributed by between-allocation variability.

Researchers can examine the extent to which PPAV differs from one parameter to another in order to judge the implications of parcel-allocation variability for inference about particular parameters of substantive interest. Note that we compute PPAV only at the final chosen M. We do not compute it iteratively at each loop in the algorithm for choosing M, nor do we use it to inform the choice of M because this practice has been critiqued in the imputation literature (e.g., Harel, 2007; Schafer, 1997). In that literature, this proportion has been shown to be noisy and imprecise at low M (Bodner, 2008; Savalei & Rhemtulla, 2012).

The second index we consider is the ratio of the between-allocation variance of a parameter estimate to the within-allocation variance (RPAV). This formula is also used in the missing data literature to compare between-imputation and within-imputation variability (Enders, 2010; Rubin, 1987; Schafer, 1997; Schafer & Olsen, 1998).

$$RPAV = \frac{V_B + (V_B/M)}{V_W}. \quad (8)$$

If the RPAV in Equation (8) is > 1, then there is relatively more parcel-allocation variability than sampling variability in that estimate; if the RPAV is < 1, then there is relatively less parcel-allocation variability than sampling variability. This index can also be calculated for each model parameter, as demonstrated below.

Demonstration of the PPAV and RPAV indices

For illustrative purposes, the PPAV and RPAV indices were calculated for each parameter from the four-factor CFA model used in Table 1 and then averaged across the 100 simulated samples. Across parameters, the PPAV ranged from .04 to .51, and RPAV ranged from .04 to 1.03. On average, across all parameters, the PPAV = .41 (i.e., on average, 41% of the total variability in an estimate is due to parcel-allocation variability), and the RPAV = .80 (i.e., parcel-allocation variability in an estimate is .80 as large as sampling variability, on average).

The PPAV for a given parameter is related to the stability of that parameter's pooled estimate and standard error across iterations of the algorithm for choosing M. In particular, a parameter with higher PPAV should require more allocations (higher M) to reach effective stability of its pooled estimate and standard error. Thus, that parameter's pooled estimate and standard error should be stable across a smaller percentage of iterations of the algorithm. This implies that a parameter with higher PPAV should individually meet the convergence criteria of the algorithm for choosing M at a smaller percentage of iterations.[6]

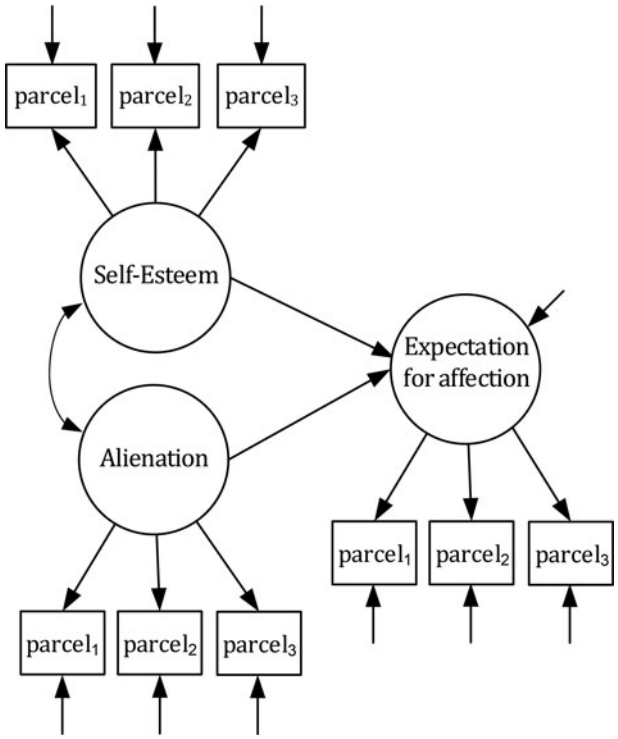

Figure 3. Path diagram for parcel-level empirical example model. Squares denote manifest indicators. Circles denote latent variables. Straight arrows denote regression paths. Curved arrows denote (co)variances.

Empirical example

We use an empirical example to illustrate the developments presented in the previous sections. Our theoretical model for this illustration is a latent variable regression model. In our model, *alienation* in the freshman year of college and *self-esteem* in the sophomore year are hypothesized to predict *expectation for affection* (i.e., expectation to be liked) in the junior year. This SEM is diagrammed in Figure 3. The path from alienation to expectation for affection is hypothesized to be negative because more prior experiences of isolation can lower expectations for peer acceptance (e.g., Lee & Robins, 1998; Williams, 2001). The path from self-esteem to expecting affection is hypothesized to be positive because youth with lower self-worth may selectively remember rejection experiences and thus have lower expectations for

[6] For example, using our four-factor simulated example from Table 2, we found a correlation of $r = -.603$ between PPAV for a parameter and the proportion of iterations (i.e., loops) of the algorithm where the convergence criteria were met for that parameter in particular. This correlation was calculated across parameters within sample and then averaged across our 100 samples.

peer acceptance (McFarlin & Blascovich, 1981). Our substantive interest lies in testing the significance of both structural paths. Self-esteem and alienation are allowed to covary.

Methods

Our empirical example employs the Jessor and Jessor (1977) Socialization of Problem Behavior in Youth: 1969–1981 study data set. In this study, a sample of $N = 205$ college students at a large state university completed surveys in the spring of each academic year. Expectation for affection is measured by ten 10-point Likert items; an illustrative item is "How important is it to be well liked by most of the people around here?" Alienation is measured by fifteen 4-point Likert items; an illustrative item is "I often find it difficult to feel involved in the things I'm doing." Self-esteem is measured by ten 4-point Likert items; an illustrative item is "How well do you make decisions about important things in your life?"

Because we are primarily interested in interpreting structural parameters in this example, we selected the number of repeated item-to-parcel allocations (M) to reach stability criteria for the structural parameters' pooled estimates and standard errors specifically. The combination scheme used for item-to-parcel allocations was as follows. Expectation for affection items were allocated into three parcels of size 3, 3, and 4 items. Alienation items were allocated into three parcels of size 5, 5, and 5 items. Self-esteem items were allocated into three parcels of size 3, 3, and 4 items.

Results

Convergence criteria of the algorithm for choosing M were reached at $M = 90$ allocations. The estimates and pooled standard errors for the paths of interest (using Rubin's [1987] rules) and the associated p values and pooled CIs were as follows. As hypothesized, an increase in freshman-year feelings of alienation significantly reduced junior-year expectation for affection, $\beta = -.213$ (.093), $p = .022$, $CI = \{-.396, -.031\}$, and an increase in sophomore-year self-esteem significantly increased junior-year expectation for affection, $\beta = .548$ (.125), $p < .001$, $CI = \{.302, .793\}$. The correlation between alienation and self-esteem was negative, $\rho = -.559$ (.096), $p < .001$. The PPAVs were .09, .28, and .27 for the three structural parameters, respectively, indicating that parcel-allocation variability contributes 9% of the variance in the effect of alienation on expecting affection, 28% of the variance in the effect of self-esteem on expecting affection, and 27% of the variance in the correlation of alienation and self-esteem. The RPAVs were .09, .38, and .38, indicating that there was over one third as much parcel allocation variability as sampling variability in some slopes. Future research could investigate whether this pattern of results can be replicated when sociometric information, instead of self-report information, is used to define alienation (e.g., Sandstrom & Cillessen, 2006).

Discussion

Parcel-allocation variability is known to arise under a variety of data and model conditions, despite having unidimensional item indicators of each factor in the population (Sterba & MacCallum, 2010). Parceling remains widely used under these conditions (e.g., modestly sized samples or communalities). Researchers may be interested in considering item-level analysis alternatives to parceling, together with categorical variable estimation methods for binary/ordinal items (Bandalos, 2008), when available and estimable. When implementing parcel-level analyses in practice, however, researchers may be interested in how to quantify and account for parcel-allocation variability. Previously, psychologists lacked information on how to: (1) combine sources of sampling variability and parcel-allocation variability when drawing inferences about parameters in SEMs, (2) choose the number of repeated allocations, and (3) quantify and report proportions of total variability per estimate due to parcel-allocating versus sampling. This article addressed these three gaps.

Regarding (1), we proposed the application of Rubin's (1987) rules to obtain pooled parameter estimates and pooled standard errors that account for uncertainty due to sampling and parcel-allocation variability within-sample. We showed that these pooled estimates and standard errors could be used for inference via NHST and CIs. Furthermore, we showed that Rubin's analytic pooling approach matches results of a Monte Carlo pooling approach implemented under the same assumptions. Regarding (2), we introduced an algorithm to select the number of allocations needed to meet a particular objective (here, that pooled estimates and standard errors retain a desired degree of stability when a specified greater number of allocations are drawn, with-replacement). We discussed alternatives for defining the convergence of this algorithm. Furthermore, we hypothesized that particular data and model conditions could increase the average number of allocations required by this algorithm, and we tested these hypotheses via simulation. We found that, indeed, under lower N, fewer items per parcel and parcels

per factor, and more factors (holding constant the numbers of items per parcel and parcels per factor), more allocations (M) were required, on average, to produce results meeting our stability criteria. Regarding (3), we provided two indices (PPAV and RPAV) to evaluate and compare the amount of parcel-allocation versus sampling variability per parameter, and we illustrated their interpretation in the context of an empirical example. In the remainder of this discussion, we address software implementation, generalizability, limitations, and future directions for our methods.

Software implementation

The algorithm for choosing M is implemented in an R function *PoolMAlloc* that is available on the authors' websites as well as incorporated into a release of the semTools *R* package.[7] A user provides an input item-level data set, a combination scheme for allocating items to parcels ($C_{q_{jk}p_k}$, see Equation [1]), values of M_{start}, M_{inc}, δ_a, and δ_b, and a specification for a parcel-level SEM, in *lavaan* (Rosseel, 2012) format. At a given iteration of the algorithm, the $M_{start} + ((h - 1) \times M_{inc})$ allocations are randomly generated[8] according to scheme $C_{q_{jk}p_k}$. Output includes the following quantities described in previous sections: the selected number of allocations M, the pooled estimates and pooled standard errors, associated p values and CI, and the PPAV and RPAV indices.[8] The reported M is the number of allocations for which model fitting was attempted in order for local stability criteria to be fulfilled using the allocation solutions that were converged and proper. Pooled estimates and standard errors are calculated across only those allocations that yielded converged and proper solutions.

Limitations and future directions

Several limitations can be noted that serve as directions for future research. First, the scope of this article was limited to pooling parameter estimates across allocations and obtaining inferences for pooled parameter estimates. Future research could consider pooling model fit statistics across allocations. Some literature has addressed this in the context of multiple imputation (Lee & Cai, 2012; Li, Meng, Raghunathan, & Rubin, 1991; Meng & Rubin, 1992), but these methods would need to be modified for the context of parcel allocations. Second, this article defined one objective for our algorithm for choosing M and developed convergence criteria that implement this objective. Future research could also consider other kinds of convergence criteria based on different objectives (see Robert & Casella, 2004) such as seeking stability in p values of parameter estimates (see Bodner, 2008; Graham et al., 2007). Third, the M obtained from this algorithm will vary across implementations of the algorithm within sample. For example, across 50 different implementations of the algorithm to the sample in Figure 1, the average chosen M was 137 with a standard deviation of 30. Researchers can rerun the algorithm multiple times within sample to get a sense of the different estimated M that can achieve similar stability in pooled results across repeated executions of the algorithm using different sets of allocations.

Tailored algorithm for choosing M versus rule-of-thumb for choosing M

In considering the results of Table 2, readers may wonder whether a fixed rule-of-thumb M that is larger than all the Ms in Table 2 (e.g., $M = 300$ or 500) could be employed in lieu of an iterative algorithm for determining M. This can be a reasonable strategy in contexts similar to those of Table 2, for example. Furthermore, computational time is shorter for a fixed rule-of-thumb approach as compared to a tailored iterative algorithm approach. For example, in the Figure 1 sample, the algorithm used on average 2.81 minutes of computational time whereas adopting a fixed number of $M = 400$ allocations used on average 0.53 minutes. Our R software also allows users to obtain pooled results for any fixed, prespecified M, and will additionally indicate elapsed computing time in the output so users can compare computing time across fixed versus automated-search approaches for determining M. Factors that can affect computing time for either approach, but are more consequential for the automated-search approach, include a greater presence of nonconverged solutions.

More broadly, however, there are two other main issues to consider when weighing these approaches. First, what may seem to be a sufficient rule-of-thumb M based on the models/data in Table 2 may be too small in other model/data contexts. For example, Sterba and Rights (in press) demonstrated that in the context of misspecification of the structural model, there is the potential for greater parcel-allocation variability, which can imply greater necessary M. The models in Table 2 have saturated structural models. If we instead misspecified the structural portion of the four-factor CFA in Table 2 such that all four factors are treated as lower-order factors loading on a

[7] See http://www.vanderbilt.edu/psychological_sciences/bio/jason-rights and http://www.vanderbilt.edu/peabody/sterba/ and https://cran.r-project.org/web/packages/semTools/index.html

[8] For details on repeatedly randomly generating allocations with replacement from $C_{q_{jk}p_k}$, see Sterba and MacCallum's (2010) *ParcelAlloc* or Quick and Schoemann's (2012) *parcelAllocation* program.

single higher-order factor (structural $df = 2$), the average number of allocations required is $M = 378$ rather than $M = 217$. Second, an advantage of employing the algorithm is that it gives the researcher empirical information about the local stability of results at the final M, whereas no such information is furnished under a fixed-rule-of-thumb approach. From a broader perspective, note that the pros and cons of a search-algorithm approach versus those of a rule-of-thumb approach are not unique to the topic of picking the number of allocations. Similarly, there are pros and cons of using such approaches in picking the number of imputations in the missing data context (Bodner, 2008; Royston, 2004). Researchers should weigh these options based on their research objectives, computing time constraints, and model/data conditions.

Generalizability

Throughout, we have suggested that our pooling approach be used when researchers are interested in using random allocations and are comfortable assuming unidimensional items on a given factor (an assumption made in most applications of parceling; Bandalos 2002, 2008; Bandalos & Finney, 2001; Hagtvet & Nasser, 2004; Hall et al., 1999; Hau & Marsh, 2004; Landis et al., 2000; Little et al., 2002; Marsh & O'Neill, 1984; Marsh et al., 2013; Matsunaga, 2008; Meade & Kroustalis, 2006; Nasser-Abu & Wisenbaker, 2006; Plummer, 2000; Rogers & Schmitt, 2004; Sass & Smith, 2006; Williams & O'Boyle, 2008; Yang et al., 2010). Prior to considering parceling, researchers may want to test unidimensionality of items on each factor—in the context of a fitted model with all factors included (Lengua, West, & Sandler, 1998)—as recommended by, for example, Marsh et al (2013) and Matsunaga (2008). There are many available sources on procedures for testing unidimensionality (e.g., Bollen, 1989; Embretson & Reise, 2000; Marsh et al., 2015; McDonald & Ho, 2002; Stucky et al., 2012). If there are known sources of multidimensionality and a researcher desires to use a purposive parceling strategy specifically designed for multidimensional items, the researcher may not want to adopt our particular pooling approach for the purposes of making inferences. In the latter setting, researchers should explicitly acknowledge in their results section that their conclusions are conditional; that is, conclusions hold for a single substantively chosen allocation, and changing the allocation could likely lead to different results (including for structural parameters). Furthermore, if there are multiple alternative ways to implement a researcher's chosen purposive strategy, the researcher should acknowledge that their reported results have not quantified uncertainty in the selection of one out of many possible allocations that could be generated by that purposive parceling strategy.

Conclusions

When investigating parcel-allocation variability in parameter estimates, previous practice had involved using an arbitrarily selected number of repeated random allocations and reporting distributions of results for each parameter. This approach did not yield a single inferential decision per parameter, making it difficult to draw substantive conclusions. To address these limitations, here we developed an algorithm for choosing the number of allocations, M, that was shown in simulations to be sensitive to model and data conditions leading to less or more parcel-allocation variability. Further, we combined sources of sampling and parcel-allocation variability to create pooled estimates and standard errors that allow a single inferential decision per parameter. The developments in this article can aid researchers in accounting for parcel-allocation variability of parcel-level results in practice. We encourage researchers to continue to consider the existence and implications of parcel-allocation variability when fitting parcel-level models.

Article information

Conflict of Interest Disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical Principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was not supported.

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The authors would like to thank Robert MacCallum, Kristopher Preacher, Keith Markus, and Sun-Joo Cho and for helpful discussions relevant to this paper. The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institution is not intended and should not be inferred.

References

Asparouhov, T., & Muthén, B. (2010a). *Plausible values for latent variables using Mplus*. Technical document available at www.statmodel.com.

Asparouhov, T., & Muthén, B. (2010b). *Multiple imputation with Mplus: Version 2*. Technical document available at www.statmodel.com.

Bagozzi, R. P., & Edwards, J. R. (1998). A general approach for representing constructs in organizational research. *Organizational Research Methods*, *1*, 45–87. doi: 10.1177/109442819800100104

Bandalos, D. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling*, *9*, 78–102. doi: 10.1207/S15328007SEM0901_5

Bandalos, D. (2008). Is parceling really necessary? A comparison of results from item parceling and categorical variable methodology. *Structural Equation Modeling*, *15*, 211–240. doi: 10.1080/10705510801922340

Bandalos, D., & Finney, S. J. (2001). Item parceling issues in structural equation modeling. In G. A. Marcoulides (Ed.). *New developments and techniques in structural equation modeling*, (pp. 269–297). Mahwah, NJ: Erlbaum.

Bayarri, M., Berger, J., Paulo, R., Sacks, J., Cafeo, J., Cavendish, J., Lin, C.-H., & Tu, J. (2007). A framework for validation of computer models. *Technometrics*, *49*, 138–154. doi: 10.1115/IMECE2003-41676

Bodner, T. E. (2008.) What improves with increased missing data imputations? *Structural Equation Modeling*, *15*, 651–675. http://dx.doi.org/10.1037/e645052007-001

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Cattell, R., & Burdsal, C. (1975). The radial parcel double factoring design: A solution to the item-vs-parcel controversy. *Multivariate Behavioral Research*, *10*, 165–179. http://dx.doi.org/10.1207/s15327906mbr1002_3

Cole, S. R., Chu, H., & Greenland, S. (2006). Multiple imputation for measurement error correction. *International Journal of Epidemiology*, *35*, 1074–1081. http://dx.doi.org/10.1093/ije/dyl097

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

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

Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, *1*, 1–32. http://dx.doi.org/10.2307/2331474

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, *7*, 457–511. http://dx.doi.org/10.1214/ss/1177011136

Ghosh-Dastidar, B., & Schafer, J. (2003). Multiple edit/multiple imputation for multivariate continuous data. *Journal of the American Statistical Association*, *98*, 807–817. http://dx.doi.org/10.1198/016214503000000738

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

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

Groves, R., & Lyberg, L. (2010). Total survey error: Past, present and future. *Public Opinion Quarterly*, *74*, 849–879. http://dx.doi.org/10.1093/poq/nfq065

Hagtvet, K. A., & Nasser, F. M. (2004). How well do item parcels represent conceptually-defined latent constructs? A two-facet approach. *Structural Equation Modeling*, *11*, 168–193. http://dx.doi.org/10.1207/s15328007sem1102_2

Hall, R., Snell, A., & Foust, M. (1999). Item parceling strategies in SEM: Investigating the subtle effects of unmodeled secondary constructs. *Organizational Research Methods*, *2*, 233–256. http://dx.doi.org/10.1177/109442819923002

Harel, O. (2007). Inferences on missing information under multiple imputation and two-stage multiple imputation. *Statistical Methodology*, *4*, 75–89. http://dx.doi.org/10.1016/j.stamet.2006.03.002

Harel, O., & Zhou, X-H. (2007). Multiple imputation: Review of theory, implementation and software. *Statistics in Medicine*, *26*, 3057–3077. http://dx.doi.org/10.1002/sim.2787

Hau, K.-T., & Marsh, H. W. (2004). The use of item parcels in structural equation modeling: Non-normal data and small sample sizes. *British Journal of Mathematical and Statistical Psychology*, *57*, 327–351. http://dx.doi.org/10.1111/j.2044-8317.2004.tb00142.x

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*, 382–417. http://dx.doi.org/10.1214/ss/1009212814

Jessor, R., & Jessor, S. L. (1977). *Problem behavior and psychosocial development: A longitudinal study of youth*. New York, NY: Academic Press.

Kish, L. (1965). *Survey sampling*. London, England: Wiley.

Landis, R. S., Beale, D. J., & Tesluk, P. E. (2000). A comparison of approaches to forming composite measures in structural equation modeling. *Organizational Research Methods*, *3*, 186–207. http://dx.doi.org/10.1177/109442810032003

Lee, R. M., & Robbins, S. B. (1998). The relationship between social connectedness and anxiety, self-esteem, and social identity. *Journal of Counseling Psychology*, *45*, 338–345. http://dx.doi.org/10.1037//0022-0167.45.3.338

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

Lengua, L. J., West, S. G., & Sandler, I. N. (1998). Temperament as a predictor of symptomatology in children: Addressing contamination of measures, *Child Development*, *69*, 164–181. http://dx.doi.org/10.2307/1132078

Li, K.-H., Meng, X.-L., Raghunathan, T. E., & Rubin, D. B. (1991). Significance levels from repeated p-values with multiply-imputed data. *Statistica Sinica*, *1*, 65–92. http://www3.stat.sinica.edu.tw/statistica/oldpdf/A1n15.pdf

Li, T. (2012). Randomization-based inference about latent variables from complex samples: The case of two-stage sampling (Unpublished dissertation). University of Maryland, College Park, MD.

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

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question,

weighing the merits. *Structural Equation Modeling*, *9*, 151–173. http://dx.doi.org/10.1207/s15328007sem0902_1

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

MacCallum, R. C. (2013). *Sells Award Address: A letter from Tuck, and how it triggered one miserable experience and then decades of research on the nature and effects of error*. St. Petersburg, FL: Annual meeting of the Society of Multivariate Experimental Psychology.

Marsh, H. W., Lüdtke, O., Nagengast, B., Morin A., & Von Davier, M. (2013). Why item parcels are (almost) never appropriate: Two wrongs do not make a right—Camouflaging misspecification with item parcels in CFA models. *Psychological Methods*, *18*, 257–284. http://dx.doi.org/10.1037/a0032773.supp

Marsh, H. W., & O'Neill, R. (1984). Self Description Questionnaire III (SDQ III): The construct validity of multidimensional self-concept ratings by late-adolescents. *Journal of Educational Measurement*, *21*, 153–174. http://dx.doi.org/10.1111/j.1745-3984.1984.tb00227.x

Matsunaga, M. (2008). Item parceling in structural equation modeling: A primer. *Communication Methods and Measures*, *2*, 260–293. http://dx.doi.org/10.1080/19312450802458935

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

McFarlin, D., & Blascovich, J. (1981). Effects of self-esteem and performance feedback on future affective preferences and cognitive expectations. *Journal of Personality and Social Psychology*, *40*, 521–531. http://dx.doi.org/10.1037/0022-3514.40.3.521

Meade, A. W., & Kroustalis, C. M. (2006). Problems with item parceling for confirmatory factor analytic tests of measurement invariance. *Organizational Research Methods*, *9*, 369–403. http://dx.doi.org/10.1177/1094428105283384

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

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*, 177–196. http://dx.doi.org/10.1007/bf02294457

Mislevy, R., Johnson, E., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*. *17*, 131–154. http://dx.doi.org/10.2307/1165155

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

Nasser-Abu, F., & Wisenbaker, J. (2006). A Monte Carlo study investigating the impact of item parceling strategies on parameter estimates and their standard errors in CFA. *Structural Equation Modeling*, *13*, 204–228. http://dx.doi.org/10.1207/s15328007sem1302_3

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Plummer, B. (2000). *To parcel or not to parcel: The effects of item parceling in confirmatory factor analysis* (Unpublished dissertation). The University of Rhode Island, Providence, RI.

Quick, C., & Schoemann, A. (2012). parcelAllocation function, in semTools R package v0.4–6.

Rässler, S. (2003). A non-iterative Bayesian approach to statistical matching. *Statistica Neerlandica*, *57*, 58–74. http://dx.doi.org/10.1111/1467-9574.00221

Ratitch, R., Lipkovich, I., & O'Kelly, M. (2013). Combining analysis results from multiply imputed categorical data (Paper SP03). *Proceedings of the Pharmaceutical Industry SAS Users Group*.

Reiter, J. (2007). *Selecting the number of imputed datasets: When using multiple imputation for missing data and disclosure limitation*. Unpublished manuscript available at: https://stat.duke.edu/~jerry/Papers/spl07.pdf

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

Robert, C., & Casella, G. (2004). *Monte Carlo statistical methods* (2nd ed.) New York, NY: Springer.

Rogers, W. M., & Schmitt, N. (2004). Parameter recovery and model fit using multidimensional composites: A comparison of four empirical parceling algorithms. *Multivariate Behavioral Research*, *39*, 379–412. http://dx.doi.org/10.1207/s15327906mbr3903_1

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

Royston, P. (2004). Multiple imputation of missing values. *The Stata Journal*, *4*, 227–241. http://www.stata-journal.com/sjpdf.html?articlenum=st0067

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

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*, 473–489. http://dx.doi.org/10.1080/01621459.1996.10476908

Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin*, *94*, 18–38. http://dx.doi.org/10.1037/0033-2909.94.1.18

Sandstrom, M., & Cillessen, A. (2006). Likeable versus popular: Distinct implications for adolescent adjustment. *International Journal of Behavioral Development*, *30*, 305–314. http://dx.doi.org/10.1177/0165025406072789

Sass, D., & Smith, P. (2006). The effects of parceling unidimensional scales on structural parameter estimates in structural equation modeling. *Structural Equation Modeling*, *13*, 566–586. http://dx.doi.org/10.1207/s15328007sem1304_4

Savalei, V., & Rhemtulla, M. (2012). On obtaining estimates of the fraction of missing information from full information maximum likelihood. *Structural Equation Modeling*, *19*, 477–494. http://dx.doi.org/10.1080/10705511.2012.687669

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

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177. http://dx.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. http://dx.doi.org/10.1207/s15327906mbr3304_5

Sterba, S. K. (2009). Alternative model-based and design-based frameworks for inference from samples to populations: From polarization to integration. *Multivariate Behavioral Research*, *44*, 711–740. http://dx.doi.org/10.1080/00273170903333574

Sterba, S. K. (2011). Implications of parcel-allocation variability for comparing fit of item-solutions and parcel-solutions. *Structural Equation Modeling*, *18*, 554–577. http://dx.doi.org/10.1080/10705511.2011.607073

Sterba, S. K., & MacCallum, R. C. (2010). Variability in parameter estimates and model fit across repeated allocations of items to parcels. *Multivariate Behavioral Research*, *45*, 322–358. http://dx.doi.org/10.1080/00273171003680302

Sterba, S. K., & Rights, J. D. (in press). Effects of parceling on model selection: Parcel-allocation variability in model ranking. *Psychological Methods*. doi: 10.1037/met0000067

Stuart, E. A., & Rubin, D. B. (2008). Matching with multiple control groups with adjustment for group differences. *Journal of Educational and Behavioral Statistics*, *33*, 279–306. http://dx.doi.org/10.3102/1076998607306078

Stucky, B., Gottfredson, N., & Panter, A. (2012) Item-level factor analysis. In H. Cooper (Ed. in Chief) *APA handbook of research methods in psychology* (pp. 683–697). Washington, DC: American Psychological Association.

Thisted, R. A. (1988). *Elements of statistical computing: Numerical computation*. London, England: Chapman & Hall.

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

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 56–75). Thousand Oaks, CA: Sage Publications.

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

Williams, K. D. (2001). *Ostracism: The power of silence*. New York, NY: Guilford Press.

Williams, L. J., & O'Boyle, E. H. (2008). Measurement models for linking latent variables and indicators: A review of human resource management research using parcels. *Human Resource Management Review*, *18*, 233–242. http://dx.doi.org/10.1016/j.hrmr.2008.07.002

Yang, C., Nay, S., & Hoyle, R. H. (2010). Three approaches to using lengthy ordinal scales in structural equation models: Parceling, latent scoring, and shortening scales. *Applied Psychological Measurement*, *34*, 122–142. http://dx.doi.org/10.1177/0146621609338592

Yuan, K.-H., Bentler, P., & Kano, Y. (1997). On averaging variables in a confirmatory factor model. *Behaviormetrika*, *24*, 71–83. http://dx.doi.org/10.2333/bhmk.24.71

Appendix: Parcel-level measurement parameters that are the subject of inference using Rubin's rules

Suppose we are interested in parcel-level measurement parameters for the *j*th parcel indicator of a particular factor. Let i denote item, s denote sample, a denote allocation, j denote the parcel of interest, and r denote the number of item indicators of that factor. Here we assume unidimensional items on a given factor; researchers typically use parceling in situations where they are comfortable assuming unidimensionality. Let λ_{is} be an item loading for item i in sample s, ε_{is} be the residual for item i in sample s, and θ_{is} be an item residual variance for item i in sample s; I_{ija} is an indicator function (0,1) determining whether item i is allocated to parcel j, in allocation a. We can also think of I_{ija} as a Bernoulli variable with parameter q_j/r, where q_j is the number of items per parcel j. $E_s()$ denotes expected value across samples; $E_a()$ denotes expected value across allocations within sample.

The factor loading for parcel j that is the subject of inference using Rubin's (1987) rules is as follows:

$$\begin{aligned} E_s\left(E_a\left(\sum_i\left(\frac{1}{q_j}I_{ija}\lambda_{is}\right)\right)\right) &= E_s\left(\frac{1}{q_j}\sum_i E_a\left(I_{ija}\lambda_{is}\right)\right) \\ &= E_s\left(\frac{1}{q_j}\sum_i \frac{q_j}{r}\lambda_{is}\right) \\ &= E_s\left(\frac{1}{r}\sum_i \lambda_{is}\right) = \frac{1}{r}\sum_i \lambda_i \end{aligned}$$

The residual variance for parcel j that is the subject of inference using Rubin's rules is as follows:

$$\begin{aligned} &E_s\left(E_a\left(\mathrm{var}_s\left(\frac{1}{q_j}\sum_i I_{ija}\varepsilon_{is}\right)\right)\right) \\ &= E_s\left(E_a\left(\frac{1}{q_j^2}\sum_i \mathrm{var}_s\left(I_{ija}\varepsilon_{is}\right)\right)\right) \\ &= E_s\left(E_a\left(\frac{1}{q_j^2}\sum_i I_{ija}^2\theta_{is}\right)\right) \\ &= E_s\left(E_a\left(\frac{1}{q_j^2}\sum_i I_{ija}\theta_{is}\right)\right) \\ &= E_s\left(\frac{1}{q_j^2}\sum_i \theta_{is}E_a\left(I_{ija}\right)\right) \\ &= E_s\left(\frac{1}{q_j^2}\sum_i \theta_{is}\frac{q_j}{r}\right) \\ &= E_s\left(\frac{1}{q_j r}\sum_i \theta_{is}\right) = \frac{\frac{1}{r}\sum_i \theta_i}{q_j} \end{aligned}$$

Note also that $E_a() = E_{allocset}\left(E_{a|allocset}()\right)$ where *allocset* refers to a set of M allocations. That is, one would get the same expectation across allocations as across allocation sets.