

Problems with Rationales for Parceling that Fail to Consider Parcel-Allocation Variability

Sonya K. Sterba

Department of Psychology and Human Development, Vanderbilt University

ABSTRACT

In structural equation modeling applications, parcels—averages or sums of subsets of item scores—are often used as indicators of latent constructs. Parcel-allocation variability (PAV) is variability in results that arises *within* sample *across* alternative item-to-parcel allocations. PAV can manifest in all results of a parcel-level model (e.g., model fit, parameter estimates, standard errors, and inferential decisions). It is a source of uncertainty in parcel-level model results that can be investigated, reported, and accounted for. Failing to do so raises representativeness and replicability concerns. However, in recent methodological literature (Cole, Perkins, & Zelkowitz, 2016; Little, Rhemtulla, Gibson, & Shoemann, 2013; Marsh, Lüdtke, Nagengast, Morin, & von Davier, 2013; Rhemtulla, 2016) parceling has been justified and recommended in several situations without quantifying or accounting for PAV. In this article, we explain and demonstrate problems with these rationales. Overall, we find that: (1) using a purposive parceling algorithm for a multidimensional construct does not avoid PAV; (2) passing a test of unidimensionality of the item-level model need not avoid PAV; and (3) a desire to improve power for detecting structural misspecification does not warrant parceling without addressing PAV; we show how to simultaneously avoid PAV and obtain even higher power by comparing item-level models differing in structural constraints. Implications for practice are discussed.

KEYWORDS

Parcel-allocation variability;
item factor analysis;
structural equation
modeling; parceling

Introduction

It is common practice for researchers to use parcel scores as indicators of latent constructs in factor analysis (FA) or structural equation modeling (SEM) applications. A *parcel* is a sum or average of a subset of item indicators of a latent construct (e.g., Cattell, 1956). In particular, between 17% and 63% of FA or SEM applications use parcel scores in lieu of item scores as indicators of latent constructs (Bandalos & Finney, 2001; Hall, Snell, & Foust, 1999; Plummer, 2000; Williams & O’Boyle, 2008). Indeed, parceling is currently considered the “prevailing approach for including scales with many items in factor analysis and SEM” (Marsh, Lüdtke, Nagengast, Morin, & Von Davier, 2013, p. 258; see also, Yang, Nay, & Hoyle, 2010, p. 123).

A *parcel allocation* is a choice of which items to allocate to a given parcel, given the researcher’s desired number of items per parcel and number of parcels per construct. *Parcel-allocation variability* (PAV) is variability in results that arises *within* sample

across alternative potential parcel allocations (Sterba & MacCallum, 2010). PAV manifests in all results of parcel-level models (i.e., in model fit statistics, model ranking, structural parameter estimates, standard errors, and inferential decisions), and the magnitude of this PAV largely depends on the amount of model error and/or sampling error (Sterba, 2011; Sterba & Rights, 2016, 2017).¹ PAV is a source of uncertainty in parcel-level model results.

PAV has implications for any individual study reporting parcel-level model results. If that individual study uses a *single* allocation, unbeknownst to the investigator results may be atypical of the distribution of results from other potential allocations within the sample. PAV also has implications across studies—involving both representativeness concerns and replicability concerns. First regarding the *representativeness concerns*: if different studies each use the *same* allocation, then their results are all conditional on the choice of that particular allocation (whereas investigators conventionally interpret results unconditionally

CONTACT Sonya K. Sterba  sonya.sterba@vanderbilt.edu  Department of Psychology and Human Development, Vanderbilt University, Peabody #552, 230 Appleton Place, Nashville, TN 37203.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hmbr.

¹Additionally, Sterba and Rights (2017) showed that even in the absence of both measurement model error and sampling error PAV in results can arise simply when there are unequal item loadings on a factor.

as if they generalize to other allocations; Maul, 2012). Second, regarding the *replicability concerns*: if different studies each use a different allocation, PAV adds to between-sample variability in results. Thus, a lack of replicability of structural results across studies may be due to substantive across-study differences (in, say, design, or sampling), or may simply be due to the use of different parcel-allocations.

In light of these representativeness and replicability concerns, existing recommendations are to quantify and report variability in parcel-level results across *repeated* allocations within a given sample (Sterba & MacCallum, 2010; Sterba & Rights, 2017). This can involve *pooling* results across allocations within sample using Rubin's (1987) rules to simultaneously account for sampling variability and parcel-allocation variability (see Sterba & Rights, 2016 for procedures). Accounting for PAV in this manner parallels how Rubin's (1987) rules are already used to simultaneously account for other non-sampling sources of variability, together with sampling variability, in a wide variety of contexts (for review, see Reiter & Raghunathan, 2007). For instance, when faced with uncertainty regarding the values missing data would take on had they been observed, researchers do not just choose one imputation of the missing data, but rather they repeatedly (i.e., multiply) impute (e.g., Enders, 2010; Little & Rubin, 2002; Rubin, 1987; van Buuren, 2012) and incorporate into the analysis the between-imputation variability in results, in addition to sampling variability. Likewise, when faced with uncertainty about the values of person-specific latent ability scores (i.e., item response theory ability scores) during secondary analyses of large scale surveys, researchers do not just choose one set of plausible values for these scores, but rather repeatedly generate sets of plausible values (e.g., Asparouhov & Muthén, 2010; Braun & von Davier, 2017; Mislevy, 1991; Schofield, Junker, Taylor, & Black, 2015; Wu, 2005) and incorporate into the analysis the between-plausible-value variability in results, in addition to sampling variability.

However, in recent methodological literature, parceling has been justified and recommended in several situations without investigating or quantifying PAV. In this article, we demonstrate problems with rationales for parceling in these situations when failing to consider PAV, and discuss implications for practice. For reference throughout the article, these situations (to be explained subsequently) are:

Situation 1: Parceling without concern for PAV when using a purposive parceling algorithm for construct(s) that are assumed to be multidimensional (Cole et al. 2016; Little et al. 2013).

Situation 2: Parceling without concern for PAV when a unidimensional-construct item-level model fits well in the sample (Marsh et al. 2013).

Situation 3: Parceling without concern for PAV when the goal is to improve power for detecting structural misspecification (Rhemtulla, 2016).

For each situation in turn, we describe the rationale provided for parceling without considering PAV, explain the problem with the rationale, provide a demonstration of this problem, and indicate consequences for applied practice. The rationales for each situation are evaluated separately, reflecting how they appeared in the literature. However, in the discussion section we contrast rationales with each other and then conclude with overall recommendations for methodological and applied research concerning parceling.

Situation 1: Parceling without concern for PAV when using a purposive parceling algorithm for construct(s) that are assumed to be multidimensional (Cole et al. 2016; Little et al. 2013)

Rationale for Situation 1

Little et al. (2013) and Cole et al. (2016) state that when substantive theory indicates that the construct of interest is multidimensional, item indicators of this construct can be allocated to parcels using a *purposive* algorithm that takes into account this multidimensionality, without a stipulation that researchers account for PAV. The implicit rationale underlying this practice is, first, that a *purposive* algorithm is a nonrandom procedure such that it would be impossible to create a within-sample distribution of results across parcel allocations generated from a given purposive algorithm, and, second, that a *random* algorithm for allocating items to parcels is inherently limited to a simple random allocating of all items to parcels with equal probability (as could be implemented for a unidimensional construct, for instance) (e.g., Little et al., 2013 p. 295, 296). As such, this rationale acknowledges that PAV can arise when using *random* allocating for a *unidimensional* construct (e.g., Little et al., 2013) but assumes it cannot arise when using a purposive algorithm for a construct theorized to be multidimensional.

Problems with the rationale for Situation 1

One problem with this rationale is that it draws too stark a distinction between random and purposive

algorithms for allocating items to parcels. In reality, popular purposive algorithms for multidimensional constructs can give rise to many possible different allocations—often random ones. As such, there can be a within-sample distribution of results across parcel allocations generated from a purposive algorithm and PAV can indeed arise when using purposive allocating.

Another problem with this rationale is that it has led simulations and empirical papers to implement purposive parceling algorithms *as if there were only one* implementation of each algorithm (i.e., as if only one possible parcel allocation could be generated from each purposive algorithm), and then to compare purposive algorithms by comparing the results of each's single allocation (e.g., Bandalos, 2008; Coffman & MacCallum, 2005; Cole et al., 2016; Hagtvvet & Nasser, 2004; Hall et al., 1999; Landis, Beal, & Tesluk, 2000; Marsh et al., 2013; Rhemtulla, 2016; Rogers & Schmitt, 2004). This practice, in turn, serves to blur and confound comparisons of purposive algorithms, by conflating *between*-purposive-algorithm variability and *within*-purposive-algorithm variability in results—as will be shown below.

Demonstration of problems with the rationale for Situation 1

To illustrate these problems, we consider what is probably the most popular purposive parceling algorithm: *heterogeneous* parceling, which is also called *domain-representative* parceling or *distributed* parceling (e.g., Hagtvvet & Nasser, 2004; Hall et al., 1999; Kim & Hagtvvet, 2003; Kishton & Widaman, 1994; Little, Cunningham, Shahar, & Widaman, 2002; Little et al., 2013; Rhemtulla, 2016). Heterogeneous parceling is applicable when the researcher's construct of interest is theorized to be multidimensional in the sense of a second-order factor. See, for example, Figure 1 Panel A. Allocating items to parcels using a *heterogeneous* parceling algorithm can entail allocating an item indicator from each lower order factor (i.e., facet) to a given parcel, as shown in Figure 1 Panel B.

According to the rationale for Situation 1, when researchers theorize that their construct of interest is multidimensional, as in Figure 1 Panel A, they can implement a heterogeneous parceling algorithm by generating a single heterogeneous parcel allocation (e.g., Figure 1 Panel B) without concern for PAV (e.g., Cole et al., 2016). However, a problem is that there is not one but rather T different ways to allocate items to parcels by applying a heterogeneous parceling

algorithm. T can be given as

$$T = (k/q)!^{q-1} \quad (1)$$

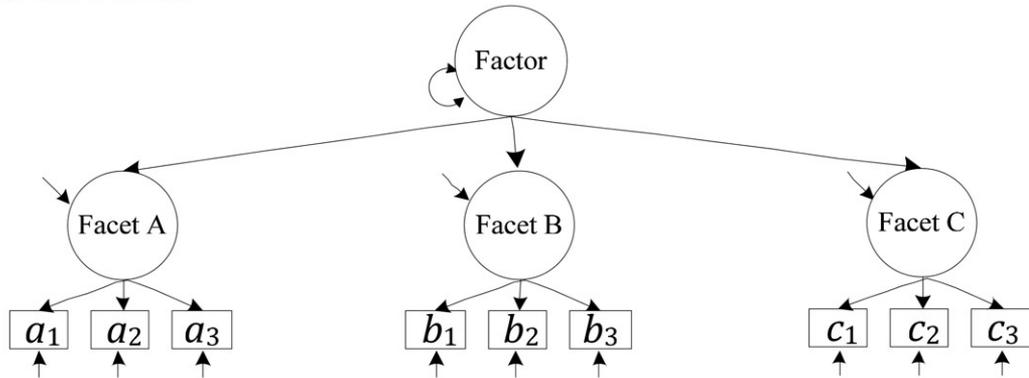
where k = the number of items and q = the number of facets, which here is also the number of items per parcel. Importantly, there can be PAV across these purposive allocations within sample. For example, consider the multidimensional extraversion construct from the Neuroticism-Extroversion-Openness Personality Inventory (NEO) (Costa & McCrae, 1985). NEO extraversion involves 36 items loading on six lower-order facets (warmth, gregariousness, excitement-seeking, etc.) such that there are six items per facet. Although NEO extraversion is often used in single-allocation heterogeneous parceling analyses (e.g., Little et al., 2002), there is not just one possible heterogeneous allocation of extraversion but rather there are 193 trillion different possible purposive allocations that implement heterogeneous parceling of extraversion.

Here, for illustration we randomly repeat this heterogeneous allocating (1 item from each lower-order facet of NEO extraversion allocated into a given parcel) $M = 1000$ times² within a sample using data from the 1987 Computer Assisted Panel Study ($N = 102$) (Latane, 1989). Within-sample PAV distributions of model fit statistics across these potential purposive heterogeneous allocations of extraversion are shown in the histograms of Figure 2 for the Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA), Standardized Root Mean Square Residual (SRMR), the p -value of the χ^2 test of absolute fit, and the p -value of the RMSEA test of close fit. Close fit is here defined as population RMSEA $\leq .05$ (Browne & Cudeck, 1993). As seen in Figure 2, the χ^2 test of absolute fit and RMSEA test of close fit both flip from significant to nonsignificant across heterogeneous purposive parceling allocations within this sample. Likewise, the CFI, TLI, and SRMR all range from good fit to poor fit across heterogeneous purposive parceling allocations within this sample. Thus, in contrast to the rationale for Situation 1, there is PAV in model fit despite having used a purposive parceling algorithm that reflects substantive theory.

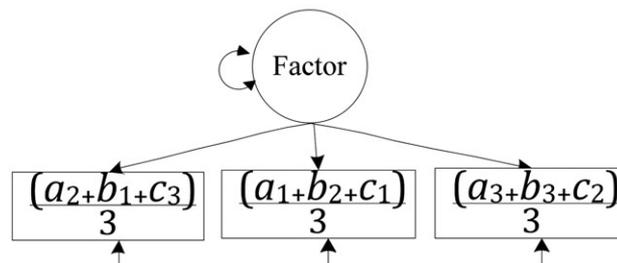
Now, suppose we wanted to compare results within-sample across different purposive parceling algorithms for our multidimensional construct of NEO extraversion. The most common comparison is between a heterogeneous parceling algorithm and a

²For discussion about choosing M see Sterba and Rights (2016).

Panel A. A theorized item-level model for a multidimensional construct (2nd-order factor). Shown with 9 items and 3 facets.



Panel B. Parcel-level model for this multidimensional construct using a single allocation from a *heterogeneous* (also called distributed uniqueness or domain representative) purposive parceling algorithm.



Panel C. Parcel-level model for this multidimensional construct using a single allocation from a *homogeneous* (also called shared uniqueness or facet representative) purposive parceling algorithm.

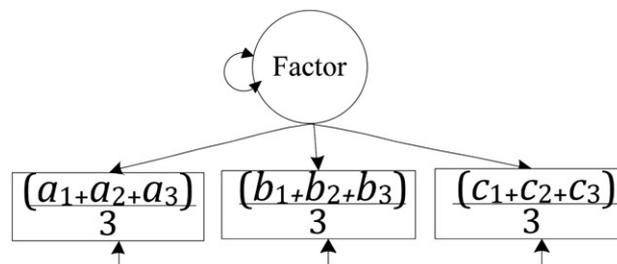
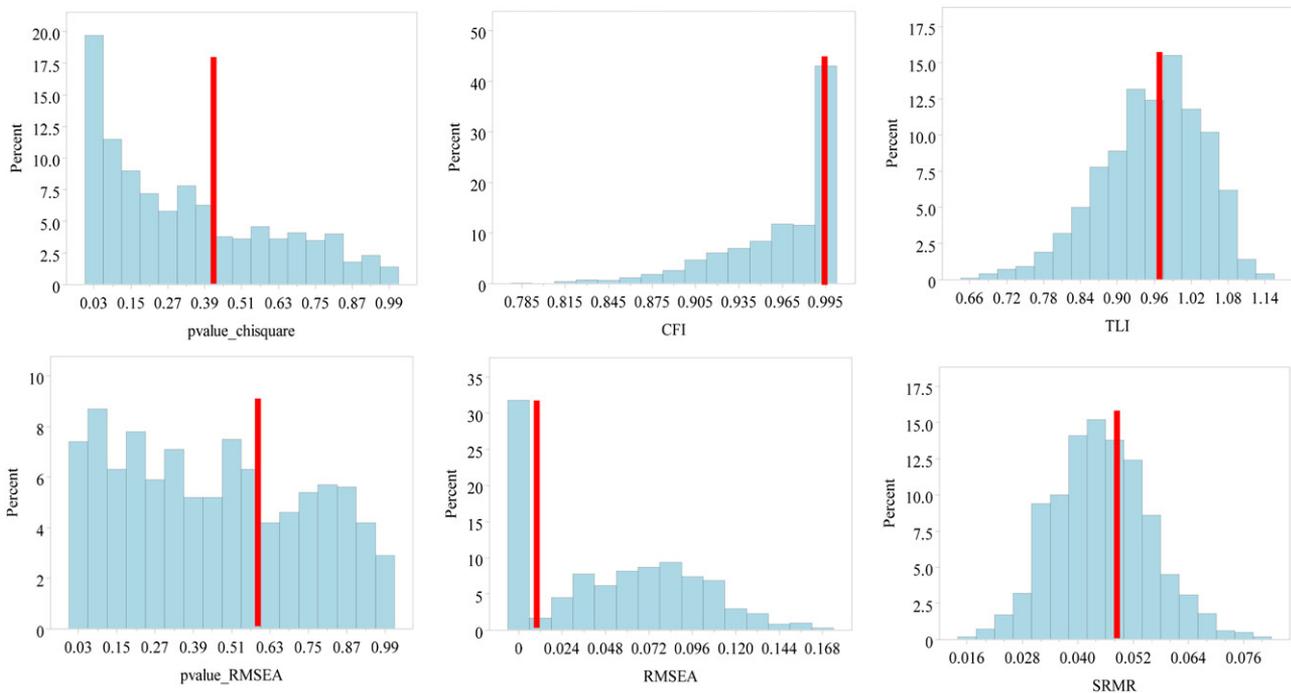


Figure 1. Background on Situation 1: Diagrams corresponding to one allocation from each of two kinds of purposive parceling algorithms (heterogeneous and homogeneous) for the same multidimensional construct.

homogeneous parceling algorithm (also called “shared uniqueness” or “facet representative” parceling). The common procedure is to use just one chosen allocation of the heterogeneous algorithm and just one chosen allocation of the homogeneous algorithm as stand-ins to represent the performance of those algorithms as a whole. The common attribution for finding differences in results from those two allocations is that the differences are reflective of the differences in the general performance of the purposive algorithms themselves (e.g., Bandalos, 2008; Coffman & MacCallum, 2005; Cole et al., 2016; Hagtvvet & Nasser,

2004; Hall et al., 1999; Landis et al., 2000; Marsh et al., 2013; Rhemtulla, 2016; Rogers & Schmitt, 2004). If we implement homogeneous parceling by allocating all item indicators of a lower-order facet to the same parcel, as shown in Figure 1 Panel C, then the model fit results from this homogeneous purposive allocation are shown by the vertical bar superimposed on each histogram in Figure 2. We can see from Figure 2 that we would draw very different conclusions about the performance of heterogeneous versus homogeneous parceling algorithms depending on *which* random allocation from the heterogeneous algorithm we chose.



Notes. *pvalue_chisquare* = *p*-value for the Chi-square test of absolute fit. CFI=Comparative Fit Index. TLI=Tucker-Lewis Index. *pvalue_RMSEA*=*p*-value for the RMSEA test of close fit. RMSEA=Root Mean Squared Error of Approximation. SRMR=Standardized Root Mean Square Residual.

Figure 2. Evaluation of Situation 1: Parcel-allocation variability in model fit for the multidimensional Extraversion construct using purposive parceling within a single empirical sample: Results from repeated *heterogeneous* parceling allocations are shown in histograms and, for comparison, results from a single homogeneous allocation is shown in the vertical bars.

That is, depending on the heterogeneous allocation, we could conclude that the homogeneous parcel-level model fit better, using any of the fit indices, or we could conclude that the homogeneous parcel-level model fit worse, using any of the fit indices. Hence, the common procedure of using just one heterogeneous allocation within-sample to represent the performance of the heterogeneous algorithm as a whole sacrifices interpretability of results because it conflates variability in results *between* purposive algorithms (homogeneous vs. heterogeneous) with variability in results *within* a given purposive algorithm.

Note that the heterogeneous- versus homogeneous-parceling comparison would become even more complex if we consider that there are also multiple ways to implement homogeneous parceling (see Rogers & Schmitt, 2004 for review); implementing these alternatives would replace the thin vertical line in each panel of Figure 2 with a second within-sample *distribution* of results in each panel of Figure 2. This second distribution would represent PAV in fit across homogeneous allocations within sample. Inferential statistics could then be employed to compare fit between homogeneous versus heterogeneous parceling algorithms by assessing whether an estimate of the

variability in fit based on its between-algorithm variability is significantly greater than an estimate of the variability in fit based on its within-algorithm PAV.

Implications for practice

Analyses using a substantively justified purposive allocation designed for multidimensional constructs (e.g., heterogeneous or distributed or domain-representative parceling) are still subject to PAV. Many simulations and empirical applications may have had different results if they had used one of the other thousands or millions of heterogeneous allocations to compare to homogeneous allocation(s). Relying on a single allocation implementation of a purposive algorithm again raises concerns about representativeness and replicability of results. Next, we turn to Situation 2.

Situation 2. Parceling without concern for PAV when a unidimensional-construct item-level model fits well in the sample (Marsh et al., 2013)

Rationale for Situation 2

Marsh et al. (2013) share the widespread methodological concern that when researchers assume a unidimensional-

construct item-level model holds based purely on substantive theory or based on previous item-level analysis from a different sample, misspecification of the item-level measurement model could be camouflaged by parceling. To allay this concern, they require a formal statistical test of the assumed item-level model to justify the use of parceling in the sample at hand. In particular, Marsh et al. (2013, p. 281) recommend that “The use of item parcels is only justified when there is good support for unidimensionality of all the constructs at the item level for the particular models and sample being considered. Tests for this requirement should be conducted for the complete model at the item level.” In other words, although Marsh et al. (2013) oppose parceling under many conditions, they nonetheless conclude that “Like Little et al., we acknowledge that the use of parcels is justified under certain circumstances, but for us these circumstances are limited primarily to models where there is support for unidimensionality at the item level” (p. 281).

The item-level testing procedure they empirically demonstrated did not additionally require statistically accounting for PAV in the parcel-level analysis after the item-level test for unidimensionality was passed. The rationale given for not addressing PAV as part of the demonstrated testing procedure was that “Parcel allocation variability will be most substantial when violations of unidimensionality are substantial” (p. 280). This would imply that PAV is larger when model error is larger. However, the relationship between PAV and model error is more complex, as will be explained shortly. Additionally, despite acknowledging that sampling error could in principle contribute to PAV, demonstrations of the effectiveness of the proposed item-level testing procedure were performed under no sampling error ($N = 100,000$), where, all else equal, the risk of PAV would be minimized. Marsh et al.’s rationale for omitting sampling error in the demonstration of the proposed testing procedure was that sampling error would cloud the picture of the test performance. Sampling error will indeed cloud this picture—but in a necessary way that is reflective of real world practice.

Problems with the rationale for Situation 2

The problem with the rationale for Situation 2 is that PAV can still arise under conditions where samples can nonetheless pass a test of unidimensionality at the item-level. Under low or no model error (i.e., low or no departure from unidimensionality in the

population) together with elevated sampling error, samples should be able to quite frequently pass (i.e., fail to reject) an item-level test of unidimensionality. Because PAV is known to occur under low or no model error plus elevated sampling error (Sterba & MacCallum, 2010), it could likewise arise under these conditions in samples passing the item-level screening test. Under moderate model error, fewer samples should be able to pass an item-level test of unidimensionality. However, PAV not only occurs in this context at either low or high sampling error, but PAV in certain results (e.g., inferential decisions) is actually maximized in this context (Sterba & Rights, 2017), and such PAV could likewise arise in samples passing the item-level screening test. Inferential decisions about the parcel-level model include decisions about rejecting the null hypothesis for tests of overall fit for the parcel-level model, tests of individual parameters in the parcel-level model, or tests comparing competing parcel-level models. Such inferential decisions are probably the most fundamental and consequential for researchers’ substantive interpretations, so PAV in such decisions is critical and concerning. The reason why PAV in inferential decisions about the parcel-level model is maximized under moderate model error is that in this setting a fit index (or test statistic) from the parcel-level model will tend to—on average, across allocations within a sample—be close to its decision threshold (or critical value), which makes it more likely for results to flip back and forth from poor fit to good fit (or significant to nonsignificant) across repeated allocations within that sample (Sterba & Rights, 2017).

Because PAV in inferential decisions about the parcel-level model is particularly elevated under moderate model error (as compared to low model error or high model error), such PAV is not monotonically related to the amount of model error. In the present context, this implies that PAV in inferential decisions is not monotonically related to the departure from unidimensionality in the population. In contrast to the rationale for Situation 2, this implies that an item-level test of unidimensionality cannot serve as a proxy test for the presence of PAV in inferential decisions about the parcel-level model. Indeed, under very high model error, virtually no samples should be able to pass an item-level test of unidimensionality, regardless of the amount of sampling error. In this situation of very high model error, PAV in inferential decisions about, for instance, parcel-level model fit should be low—because fit indices would (on average across

allocations within a sample) be far from their decision threshold (Sterba & Rights, 2017)—though PAV in other results (e.g., PAV in point estimates) could continue to increase.

In sum, though well-intentioned, if adopted in practice in its demonstrated form, an item-level testing procedure for unidimensionality risks giving researchers false assurance that they can ignore PAV in subsequent parcel-level analyses, under circumstances where PAV can nonetheless be large enough to affect substantive conclusions.

Demonstration of problems with the rationale for Situation 2

Consider the scenario where a researcher theorizes that he or she has three correlated unidimensional constructs. Typical practice would be to rely on previous results and/or substantive theory regarding the factor structure, and then parcel item indicators within each factor, and exclusively fit a 3-factor parcel-level model. Here, we evaluate the strategy proposed in Situation 2 of *testing* the unidimensionality of the 3-factor item-level model prior to parceling. We use a simulation to investigate whether this test serves as an indicator of the presence of PAV—such that passing the test of unidimensionality implies no or nonmeaningful PAV.

Before continuing, it is worth noting that Situation 2 reverses the logic commonly employed in applied practice. Currently researchers often decide to parcel when they obtain a *poorly fitting* item-level model or when they *cannot* estimate an item-level model (Bagozzi & Edwards, 1998; Hau & Marsh, 2004; Irwing, Booth, & Batey, 2014; Little et al., 2013; Martens, 2005; Matsunaga, 2008; Meade & Kroustalis, 2006; Nasser-Abu & Wisenbaker, 2003; Plummer, 2000; Sainio et al., 2013; Williams & O’Boyle, 2008; Yang et al., 2010). On the contrary, Marsh et al. (2013) is saying that researchers *must obtain a well-fitting* unidimensional construct item-level model *to parcel*.³ Hence, under Situation 2, there would be many samples in which applied researchers simply would not be allowed to parcel—either because the item-level model does not fit adequately or the item-level model is not computationally feasible. Here, we do not focus on those samples (nor what analysis options would be left open for them). Rather, we

³One may wonder why researchers would want or need to parcel if they can estimate an item-level model, but we do not take up this matter here (see Bandalos, 2008, and see Situation 3 and its response below) because our purpose here is to evaluate the implications of Situation 2 for PAV.

focus on those samples that *would* pass the test, indicating a well-fitting unidimensional-construct item-level model.

Simulation design

The simulation design involves 20 cells, with 500 samples generated per cell. There are four sample sizes: 100, 200, 300, and 500 (manipulating sampling error) and five generating item-level models (manipulating measurement model error). Each item-level generating model is a confirmatory factor analysis (CFA) model with three primary factors, nine items per factor, item loadings alternating among 0.4, 0.5, and 0.6 per factor, factor variances of 1.0, correlations among the three primary factors of 0.25, 0.50, and 0.25, and item residual variances chosen to make all items have unit variances.

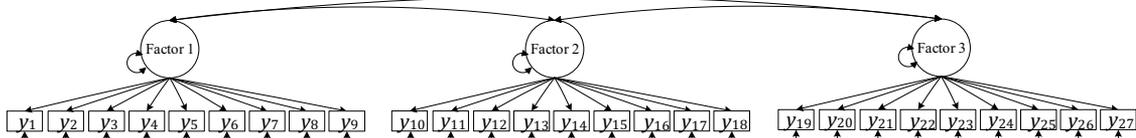
Item-level generating models differed in the extent to which they depart from construct unidimensionality. That is, generating models differed in the presence/absence of error covariances and a method factor, as shown in Figure 3. In generating Model A, items in the population truly are unidimensional per factor in the population, which corresponds to the substantively theorized model. In Generating Model B the departure from three unidimensional constructs is small, corresponding to an RMSEA of .018 in the population; this model is similar to those used in previous parceling literature simulations (e.g., Bandalos, 2002, 2008). It has a method factor on which five items from factor 1 load and 5 items from factor 2 load, with loadings of 0.3. In Generating Model C the departure from 3 unidimensional constructs is medium, corresponding to an RMSEA of .055 in the population; it adds eight error covariances of size 0.25 to Model B. In Generating Models D and E the departure from three unidimensional constructs is medium-large or large—corresponding to an RMSEA of .063 or .081 in the population, respectively. Generating Models D and E add, respectively, 3 or 6 more error covariances of size 0.25 to Model C. Generating Models A–E are diagrammed in Figure 3.

Simulation analysis overview

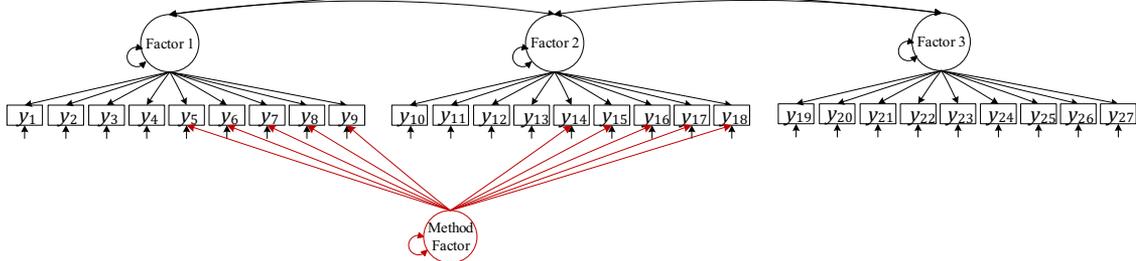
A flowchart overview of the simulation analysis procedure is given in Figure 4. First, the theorized item-level model (three correlated unidimensional constructs) is tested⁴ in each sample. If that test is *not*

⁴The test used will be described shortly.

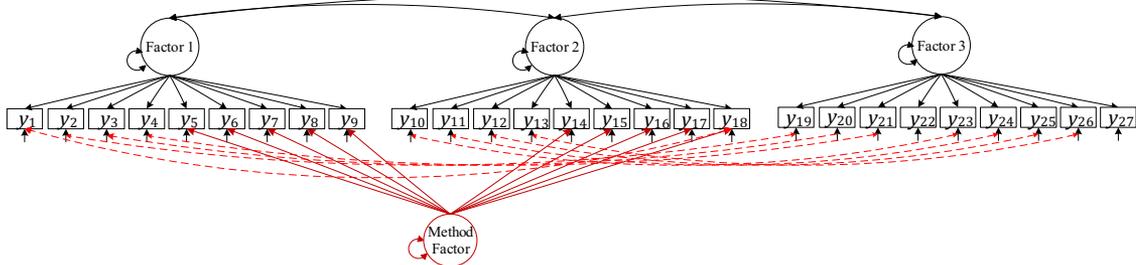
Generating Model A: 3-factor model with 3 unidimensional constructs (researcher’s theorized model).



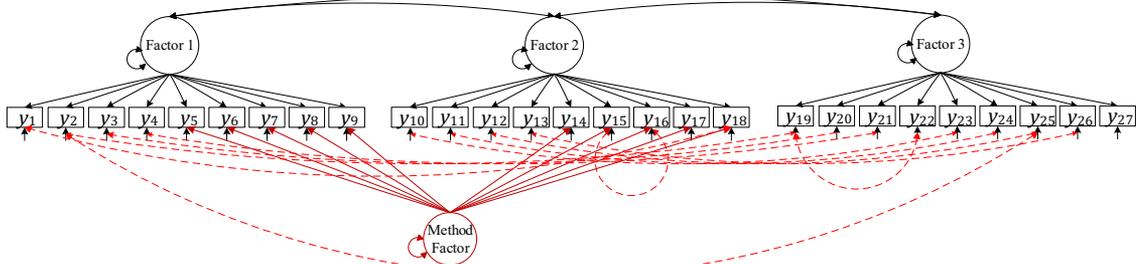
Generating Model B: Departure from 3 unidimensional constructs corresponds to RMSEA=.018 in population. This generating model is similar to that used in previous parceling literature (e.g., Bandalos, 2002, 2008).



Generating Model C: Departure from 3 unidimensional constructs corresponds to RMSEA=.055 in population.



Generating Model D: Departure from 3 unidimensional constructs corresponds to RMSEA=.063 in population.



Generating Model E: Departure from 3 unidimensional constructs corresponds to RMSEA=.081 in population.

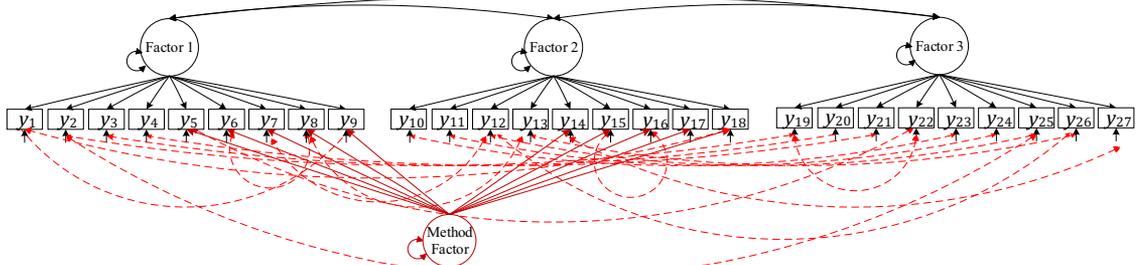


Figure 3. Generating item-level models used for the simulation evaluating Situation 2.

passed in a given sample, parceling is not permitted under Situation 2 and this sample is not analyzed further here. However, if that test is passed in a given sample, we can parcel under Situation 2. So in each sample where this test was passed, we randomly

allocate items to parcels within factor 500 different times and then we fit the 3-factor parcel-level model (see Figure 4) to each of these 500 allocations within sample. Finally, we assess PAV in parcel-level model results within that sample.

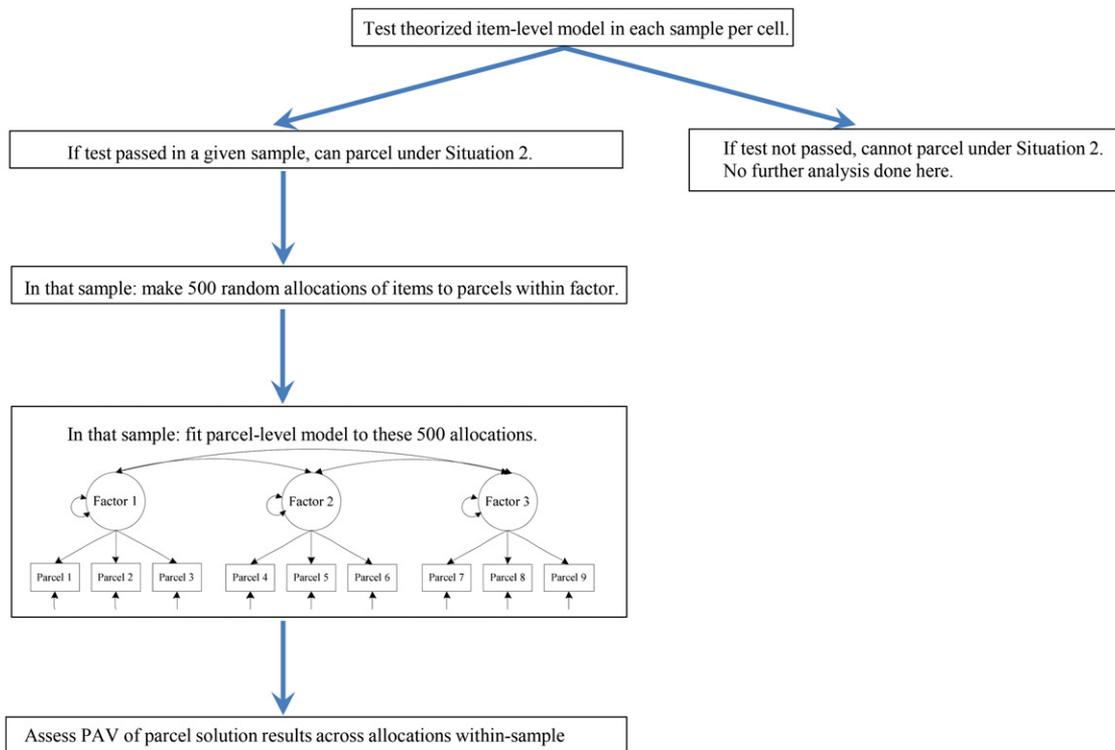


Figure 4. Analysis flowchart for the simulation evaluating Situation 2.

Step 1: Test of unidimensionality for the item-level model

Regarding the testing procedure for the item-level model, there are many published approaches for testing construct unidimensionality (e.g., Bandalos, 2002; Bonifay, Reise, Scheines, & Meijer, 2015; Finch & Habing, 2007; Gerbing & Anderson, 1988; Hagtvet & Nasser, 2004; Hall et al., 1999; Hattie, 1985; Hoskens & De Boeck, 1997; Matsunaga, 2008; Plummer, 2000; Reise, Scheines, Widaman, & Haviland, 2013; Stucky, Gottfredson, & Panter, 2012; West, Finch, & Curran, 1995). Although the item-level model is rarely tested in empirical parceling applications (as noted in Bandalos & Finney, 2001; Marsh et al., 2013), the method most commonly mentioned for doing so in the parceling literature is to test the absolute fit of an item-level CFA model with unidimensional constructs and conclude that the test is passed if, for instance, close fit (here defined as population $RMSEA \leq .05$) (Browne & Cudeck, 1993) cannot be rejected using the RMSEA. Employing exploratory factor analysis (EFA) is an alternative that has been mentioned less frequently in the parceling literature. Marsh et al. (2013) suggested a more stringent testing approach that is employed here for evaluating Situation 2. This approach involves: testing the absolute fit of the theorized 3-factor item-level CFA, and

then also fitting an EFA with the same number of factors (here, 3). Parceling is admissible according to this testing approach if (a) the item-level CFA fits well, and also (b) the fit of the item-level EFA is not significantly better than that of the item-level CFA. Note that Marsh et al. (2013) additionally preferred that the structural covariances would not meaningfully differ between the EFA and CFA solutions as part of this testing procedure. However, they neither operationalized what would constitute large versus small differences, nor provided a formalized statistical and objective method for their comparison (they gave only single-sample demonstrations where differences were informally visually inspected); hence we cannot implement their additional suggestion here. We later address the implications of this exclusion in the “Implications and future directions” portion of this section.

To mirror the most common empirical practice, we consider the test passed if (a) close fit of the item-level CFA model cannot be rejected using RMSEA, and (b) there is a nonsignificant likelihood ratio difference test between the item-level CFA and item-level EFA.

Simulation results for testing the item-level model are as follows. Under the correct specification (Generating Model A), we are allowed to parcel (i.e.,

the item-level test of unidimensionality is passed) in 86%, 91%, 92%, and 94% of samples across N 's of 100, 200, 300, and 500, respectively. Under the smallest misspecification (Generating Model B), we are allowed to parcel in 80%, 71%, 62%, and 42% of samples across N 's of 100, 200, 300, and 500, respectively. Under medium misspecification (Generating Model C), we are allowed to parcel in 30%, 24%, 12%, and 1% of samples across N 's of 100, 200, 300, and 500. Under medium-large misspecification (Generating Model D), we are allowed to parcel in 11%, 5%, 1%, and 0% of samples across N 's of 100, 200, 300, and 500, and under the largest misspecification, we can parcel in 0% of samples across N .

Step 2. Parcel-analysis for samples passing the test of unidimensionality for the item-level model

Next, we separated the samples in each cell where unidimensionality was versus was not supported in item-level analysis. In each of the samples where unidimensionality was supported in item-level analyses (i.e., where the test of unidimensionality was passed at the item-level in Step 1), parcels were formed by randomly allocating items to parcels within factor 500 different times per sample. This yielded 500 parcel-level data sets in each sample. As mentioned earlier, each of these parcel-level data sets was then fit with a 3-factor parcel-level CFA. Among these samples passing the stringent test of unidimensionality permitting parceling, we found that PAV still arises, in contrast to the rationale for Situation 2. Although PAV can be quantified in any kind of parcel-level model result, here for illustration we show PAV in absolute fit of the parcel-level model using RMSEA. A within-sample PAV distribution of parcel-solution RMSEAs exists for every sample passing the test of unidimensionality, and in theory each of these distributions could be plotted. To reduce the number of figures, we instead plot *pooled* within-sample across-allocation distributions of RMSEA, for each cell, in Figure 5.⁵

In Figure 5, Panels A–D correspond to generating Models A–D and each distribution in a given panel

corresponds to a different sample size. Note that in Figure 5 each distribution appears bimodal simply because the RMSEA fit index is bounded from below by zero. Observe that these within-sample PAV distributions of parcel-solution RMSEAs span the commonly used cutoff of .05, indicating that fit is flipping from good to poor across randomly drawn parcel-allocations within samples that nonetheless passed the stringent test of unidimensionality.

To complement the graphical results in Figure 5, for each allocation in each of these samples, PAV in the inferential decision about the parcel solution's RMSEA test of close fit was recorded. These results indicated that, when there is no or small measurement model misspecification (Figure 5 Panels A and B) at lower N at least a fourth of these samples (26–37%) exhibit PAV in the significance of the test of close fit for parcel-solutions within-sample. And for medium or medium-large measurement model misspecification (Panels C and D) across all N 's, >91% of these samples (Panel C) or across all N 's >83% of these samples (Panel D) exhibit PAV in the significance of the test of close fit for parcel-solutions within-sample.

We now summarize the overall pattern of results from the simulation analysis Steps 1 and 2, taken together. Under low or no model error, more samples pass the item-level test of unidimensionality, but a lower proportion of these samples have PAV in the parcel-level test of close fit. In contrast, under moderate model error, fewer samples pass the test of unidimensionality, but the great majority of these samples have PAV. Only with a large amount of model error do no samples pass the test of unidimensionality, meaning that parceling is not allowed according to Marsh et al. (2013); hence PAV was not investigated here in this condition.

Implications for practice

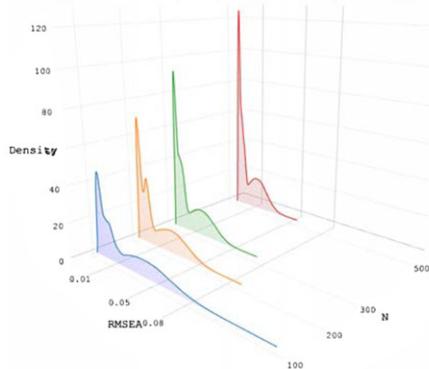
This simulation demonstration showed that a test of fit of a unidimensional-construct item-level model (even if there are indeed unidimensional constructs in the population) does not effectively function as a test of when a researcher may parcel without investigating PAV. In samples that pass such a test, PAV can still arise. In samples with moderate model error that pass such a test, there are especially high rates of PAV.

The solution is not that we need to, say, try a different test of unidimensionality of the item-level model. If we had used a *less stringent* test of

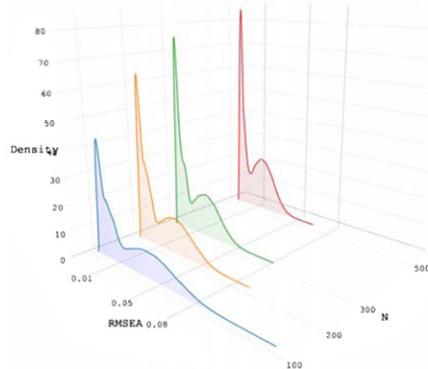
⁵The pooled PAV distribution for samples that passed the item-level test in a given cell of the simulation is computed in the following manner. Take each within-sample across-allocation distribution of parcel-solution RMSEAs and subtract the sample mean of the parcel-solution RMSEAs. Then, pool results across those samples, yielding a distribution of sample-mean-centered parcel-solution RMSEAs which is essentially a pooled-within distribution of parcel-solution RMSEAs. Then, add back in the cell-mean of the parcel-solution RMSEAs.

Panel A. Correct specification condition:

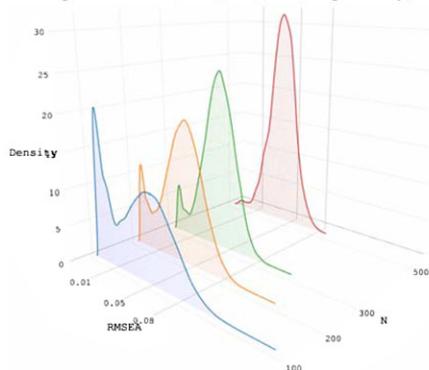
Only samples passing item-level test of unidimensionality contribute to these plots (86%, 91%, 92%, & 94% of samples at N=100, 200, 300, 500, respectively)

**Panel B. Small misspecification condition:**

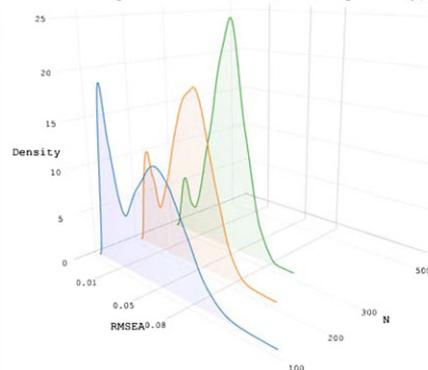
Only samples passing item-level test of unidimensionality contribute to these plots (80%, 71%, 62%, & 42% of samples at N=100, 200, 300, 500, respectively)

**Panel C. Medium misspecification condition:**

Only samples passing item-level test of unidimensionality contribute to these plots (30%, 24%, 12%, & 1% of samples at N=100, 200, 300, 500, respectively)

**Panel D. Medium-large misspecification condition:**

Only samples passing item-level test of unidimensionality contribute to these plots (11%, 5%, and 1% of samples at N=100, 200, 300, respectively)



Note. Each distribution in Figure 5 is a pooled within-sample across-allocation distribution of a parcel-solution fit index (Root Mean Squared Error of Approximation, RMSEA), constructed for samples that passed the item-level test of unidimensionality in a given cell of the simulation. Each distribution is obtained in the following manner. Take each within-sample across-allocation distribution of parcel-solution RMSEAs and subtract the sample-mean of the parcel-solution RMSEAs. Then pool results across those samples, yielding a distribution of sample-mean-centered parcel-solution RMSEAs. Then add the cell-mean of the parcel-solution RMSEAs back in. Plot.ly (3d scatter command) in R was used in creating this figure (Plotly Technologies, Inc., 2015). Note that the distributions appear bimodal because the RMSEA fit index is bounded from below by zero. This figure appears in color in the online, but not print, version of the article.

Figure 5. Evaluation of Situation 2: Parcel-allocation variability (PAV) in parcel-solution model fit for samples passing a test of unidimensionality of the item-level model.

unidimensionality of the item-level model, then we would again have PAV to contend with. If we had used a *more stringent* test of unidimensionality of the item-level model, as Marsh et al. (2013) would prefer,⁶ then parceling would have been allowed in fewer and fewer samples. But among those few samples where parceling is still allowed, PAV will still arise.

⁶As mentioned earlier, Marsh et al. (2013) preferred a final step to make their testing procedure even more stringent: inspecting structural covariances to see if they meaningfully differ between the EFA and CFA solutions. But they did not supply an objective way to operationalize this aspect of the procedure, and so it was not implemented here.

It is also possible that, under different data generating conditions, applying the same test of unidimensionality used here could allow parceling in even fewer samples because of differential sensitivity of EFA and CFA to certain kinds of departures from unidimensional items (e.g., cross loadings). But again, among the remaining samples where parceling is allowed by the testing procedure, PAV would still arise.

The bottom line is that a test of unidimensionality of the item-level model is not an effective indicator of whether the magnitude of PAV could substantively

alter study results. It is safer to investigate and quantify PAV in the sample at hand when intending to parcel. Next, we turn to Situation 3.

Situation 3. Parceling without concern for PAV when the goal is to improve power for detecting structural misspecification (Rhemtulla, 2016)

Rationale for Situation 3

To date, parceling has been widely discouraged when research interest lies in scale development or refining the measurement model (Rhemtulla, 2016); instead, when researchers implement parceling they most typically state that their interest lies in structural relations among factors, under the assumption of a well-understood, known population item-level measurement model (e.g., Bandalos, 2002; Bandalos & Finney, 2001; Little et al., 2002, 2013; Marsh et al., 2013; Matsunaga, 2008; Meade & Kroustalis, 2006; Nasser-Abu & Wisenbaker, 2006; Plummer, 2000; Rogers & Schmitt, 2004; Sass & Smith, 2006; Stucky et al., 2012; Williams & O’Boyle, 2008). In this commonly encountered context where researchers are willing to assume a correctly specified measurement model, Rhemtulla (2016) further recommends using parceling to improve power for detecting structural misspecification, stating “Most crucially, parcels vastly improve power to detect misspecification in the structural model. The goal of most research using SEM is to test the structural relations among constructs. Parcels help make that goal more attainable” (p. 366). However, this recommendation did not include a stipulation that researchers account for PAV. The rationale for this recommendation was framed in terms of pros and cons of parceling. Improved power for structural parameter tests was viewed as a motivation for parceling, to be weighed against the drawback of PAV—the implication being that if the former was compelling enough, perhaps parceling would still be warranted even if PAV were ignored.

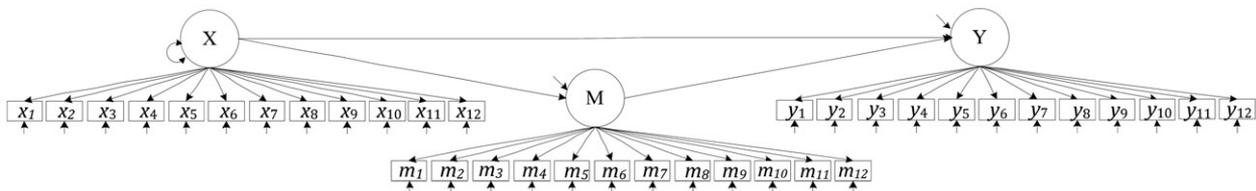
Problems with the rationale for Situation 3

There are three problems with the rationale underlying Situation 3. The first problem is that, as evidenced in the simulation from the Situation 2 section, PAV arises under the very condition assumed by Situation 3 (i.e., a correctly specified measurement model) and moreover arises even when the item-level model fits adequately in the sample at

hand. So it is not justifiable to parcel without concern for PAV in this context—even if parceling were to pose a benefit in terms of improving power for detecting structural misspecification. An implication for applied researchers is that they should not conceptually tally pros and cons of parceling and perceive that its beneficial and detrimental effects can effectively cancel out in the grand scheme; benefits of parceling do not stand on their own to justify ignoring PAV.

The second problem with the rationale underlying Situation 3 is that the gain in power for the parcel-level model was shown only theoretically in Rhemtulla (2016) (i.e., using theoretical power, defined shortly), but it does not hold empirically under commonly employed circumstances (i.e., using empirical power, defined shortly). Instead, under commonly employed empirical circumstances the opposite pattern holds: item-level power is greater than parcel-level power for the test Rhemtulla (2016) used (a global absolute test of fit of the structural-plus-measurement model under the assumption that the measurement portion is correctly specified). Therefore Rhemtulla’s theoretical power results are not reflective of the item-versus-parcel power difference that actually manifests in empirical practice when parceling is used to detect structural misspecification with this test. This is only part of the issue, however, because of course, empirical power results could be analytically adjusted to conform more with theoretical results, though this is not done in applied practice and is not incorporated in SEM software.⁷ The other part of the issue—our third problem with the rationale—is that the test that was used in Rhemtulla (2016) for detecting structural misspecification (a global test of absolute fit of the measurement-plus-structural model, under the assumption that the measurement portion is correctly specified) is insensitive for the stated purpose of detecting structural misspecification. It will be shown here that this test provides much lower power—whether using theoretical power *or* empirical power and whether using an item-solution *or* a parcel-solution—as compared to a model comparison approach that compares models differing in structural specification (as in Sterba & Rights, 2017). Moreover, it will be shown here that Sterba and Rights’ (2017) structural model comparison approach not only can yield higher power, but moreover can yield approximately equal power for the item-solution and parcel-solution, regardless of whether theoretical power or empirical

⁷Sterba and Preacher (in prep) discuss how and when to make such adjustments to the χ^2 test of absolute fit.



Notes. Fitted models (described in Figures 7 and 8) evaluate the misspecification of omitting the direct effect of the x -factor on the y -factor in Figure 6.

Figure 6. Generating model from Rhemtulla (2016) used in evaluating Situation 3.

power calculations are used. This equivalency result undermines Rhemtulla's (2016) recommendation to parcel to achieve higher power for testing structural constraints, as parceling is not then necessary for this purpose. Moreover, parceling without accounting for PAV in this context then lacks a justification. These second and third problems are demonstrated in the next section.

Demonstration of problems with the rationale for Situation 3

We begin by reproducing the evidence Rhemtulla (2016) supplied to show that parceling improves power for detecting structural misspecifications.

Theoretical power for a global absolute test of fit of the structurally constrained model

The method Rhemtulla (2016) used to compute power was theoretical—that of Satorra and Saris (1985)—and the test for which power was computed was a χ^2 test of global absolute fit of the full (measurement-plus-structural) model with structural constraints—assuming that the measurement portion was correctly specified. Her generating item-level latent variable mediation model is shown in Figure 6 and fitted (item-level and parcel-level) models are shown in Figure 7.

As can be seen from Figure 7, the fitted models had misspecification only in the structural portion (an omitted direct path). This example will serve as our Situation 3 running example. In this example, the largest discrepancy in item-level versus parcel-level power was at a modest N around 150, wherein analytic results for the item-level $\chi^2(592)$ test of absolute fit here yielded power=.08 but the power for the parcel-level $\chi^2(25)$ test of absolute fit was indeed higher:

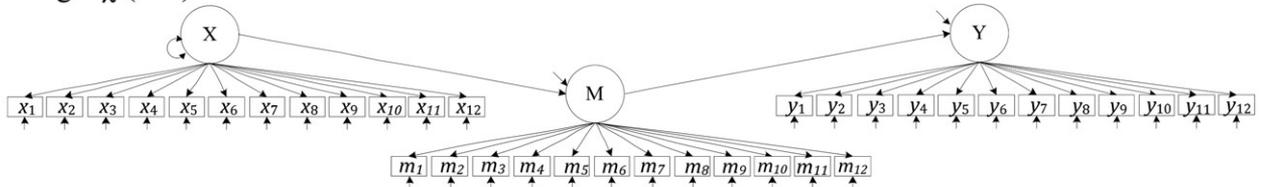
power = 0.28.⁸ Why is there greater theoretical power at the parcel-level than item-level here? The reason is that there are many more degrees of freedom for the item-level model ($df = 592$) than the parcel-level model ($df = 25$) but there is nearly identical noncentrality for the item-level model (noncentrality = 8) and parcel-level model (noncentrality = 8) because the misspecification arises purely from the structural portion of the model (relatedly, see Mulaik et al., 1989).

Empirical power for a global absolute test of fit of the structurally-constrained model

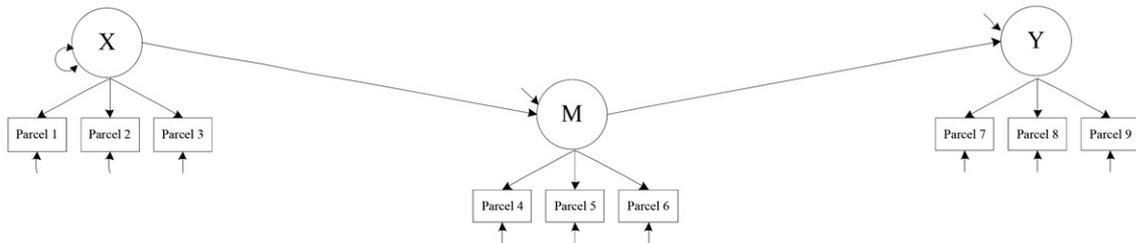
In contrast to the theoretical power results supplied above, we next demonstrate that, in conditions encountered in common empirical practice, parceling does not improve power, vis-a-vis the item-solution, for detecting structural misspecification using this global test of absolute fit. Common empirical practice for parceling applications is reflected by sample sizes below 1000 but a large number of items (p). However, it is under this very situation that the χ^2 statistic is inflated and does not follow its theoretical distribution well (e.g., Shi, Lee, & Terry, 2018; Yuan, Tian, & Yanagihara, 2015). Hence, in parceling applications *empirical power*, but not theoretical power, will mirror the actual magnitude of power that an applied researcher could expect to have in practice when fitting a misspecified model to a real dataset using standard SEM software and computing this χ^2 test of

⁸The same pattern of results held for Rhemtulla's generating parameters from her Model 3b. Here, we used a similar set of generating parameters: standardized latent regression paths were 0.4 except for the direct effect of the x -factor on the y -factor, which was 0.35; standardized item loadings were 0.4. Note that we used these generating parameters because the even larger effect sizes in the original source led to a ceiling on power at 1.0 when we implemented our alternative testing approach (described subsequently) but not when implementing the originally reported approach, which made our

Panel I. Fitted item-level model used to compute power for detecting the structural misspecification using a $\chi^2(592)$ test of absolute fit.



Panel II. Fitted parcel-level model used to compute power for detecting the structural misspecification using a $\chi^2(25)$ test of absolute fit.



Notes. The misspecification was only in the structural portion (an omitted direct effect of the x-factor on the y-factor).

Figure 7. Background on Situation 3: To compute item-level and parcel-level power for detecting a structural misspecification, Rhemtulla (2016) fit the structurally misspecified model as an item-level model (Panel I) or parcel-level model (Panel II) and assessed power for the global absolute test of fit.

absolute fit.⁹ Theoretical power for the χ^2 test of absolute fit uses a theoretical null distribution and a theoretical alternative distribution (e.g., Satorra & Saris, 1985). Empirical power for the χ^2 test of absolute fit uses a theoretical null distribution but an empirically generated alternative distribution, obtained via repeated sampling; empirical power is the area under this empirically generated alternative distribution beyond the critical value defined under the theoretical null distribution.¹⁰ Empirical power is computed by generating repeated samples (here 5000) from a model and then, to each sample, fitting the structurally misspecified model and computing the χ^2 test of absolute

fit using standard SEM software, and then calculating the proportion of tests that are significant. For the same structural misspecification and same sample size ($N=150$) used above, empirical power shows the *opposite* pattern than was found theoretically by Rhemtulla (2016)—for the item-level analysis, empirical power is 0.63, double that of the parcel-level analysis, for which empirical power is 0.29. This empirical approach—unlike the theoretical approach—also allows us to acknowledge and report the amount of PAV in power that would manifest in practice; here, for a parcel-level analysis, power for the absolute test of fit ranged from 0.27 to 0.33 depending on the allocation.¹¹

This higher empirical power for detecting structural misspecification in the item-level model contributes to what practitioners have encountered in practice in the typical situation of moderate N and large p . In practice—where a common objective of a parcel-analysis is

⁹The same logic applies to, for instance, RMSEA, which is a function of the χ^2 , but we do not repeat our demonstration for multiple fit indices here.

¹⁰Note that what we term empirical power for the χ^2 test of absolute fit differs from what Yuan, Zhang, and Zhao (2017) term Monte Carlo power. As explained and illustrated in detail in Sterba and Preacher (in prep), Yuan et al. (2017) use both an empirically generated null distribution and an empirically generated alternative distribution. Power computed using their approach will not mirror the actual power that manifests in real-world practice with low N and high p after researchers collect data and fit their model using standard SEM software. (Standard SEM software by default uses a theoretical null distribution, not an empirically generated null distribution, when performing a χ^2 test of absolute fit.)

¹¹Although our focus here is not on PAV in structural parameter estimates, note that in this simulation the average across-allocation within-sample range for the point estimate of the direct effect of the x-factor on the y-factor (c-path) was {0.28–0.43} and the average across-allocation within-sample range for the standard error of the direct effect (c-path) was {0.11–0.15}.

to test constraints on a structural model which is likely somewhat misspecified—researchers frequently comment that their item-solution “fits worse” and their parcel-solution “fits better” as well as frequently comment that their item-solution more often rejects H_0 for the global χ^2 test of absolute fit (e.g., Bagozzi & Edwards, 1998; Bagozzi & Heatherton, 1994; Bandalos, 2002; Gribbons & Hocevar, 1998; Hagtvet & Nasser, 2004; Landis et al., 2000; Little et al., 2002; Marsh et al., 2013; Martens, 2005; Matsunaga, 2008; Nasser & Wisenbaker, 2003; Plummer, 2000; Rogers & Schmitt, 2004; Schallow, 2000; Sterba, 2011; Takahashi & Nasser, 1996; Thompson & Melancon, 1996; Williams & O’Boyle, 2008). Such comments are another way for researchers to say that their item-solutions had higher empirical power in practice than their parcel-solutions.

In sum, the above demonstration communicates that the theoretical advantage in power from Rhemtulla (2016) using parceling together with a global absolute test of fit to detect structural misspecification is likely not borne out in empirical practice. Fortunately, to detect structural misspecification, we do not need to choose between using a global test of absolute fit with a parcel-solution (with better theoretical power and worse empirical power) versus an item-solution (having better empirical power and worse theoretical power).¹² The reasons are that (a) the global test of absolute fit itself is insensitive for the stated purpose of detecting structural misspecification, and (b) an alternative testing approach will yield much *higher power* for detecting the structural misspecification, and it will do so regardless of whether an item-level or parcel-level analysis is used and regardless of whether power is computed theoretically or empirically. Moreover, in this case the alternative testing approach shows virtually no discrepancy between item-solution and parcel-solution power, regardless of whether power is computed theoretically or empirically. This means that it is not in fact necessary to use parceling to improve power for detecting structural misspecifications, as was stated in Rationale 3.

¹²Such a choice would be fraught. Although the parcel-solution has a theoretical power advantage using the global test of absolute fit for detecting structural misspecifications, it has a disadvantage for detecting measurement model misspecification (e.g., Bandalos, 2002; Hall et al., 1999; Meade & Kroustalis, 2006; Rhemtulla, 2016), and in the likely context where at least a little of both kinds of misspecification co-occur, such advantages and disadvantages could cancel out. Furthermore, although the item-solution has an empirical power advantage using the global test of absolute fit for detecting structural misspecification in the common circumstances of moderate N and large p , it also has elevated type I error under these circumstances [unless adjustments to the χ^2 test of absolute fit are made to allow empirical type I error and power to more closely conform with theoretical expectation—see Sterba and Preacher (in prep) for procedures].

Empirical and theoretical power for detecting structural misspecification by comparing structurally-constrained and structurally-unconstrained models

This alternative approach to detecting structural misspecification (again under the same assumption of a correctly specified measurement model), involves comparing the fit of competing models differing in their structural constraints (e.g., Sterba & Rights, 2017). This could be done with model selection indices, but here is shown with a χ^2 difference test. For our Situation 3 running example, this structural model comparison approach is diagramed in Figure 8 and involves comparing the fit of models with (Model B) and without (Model A) the structural direct path between the X -factor and Y -factor. Next we demonstrate and explain the aforementioned features of this approach.

Figure 9 shows that theoretical and empirical power computations agree¹³ that, when using a structural model comparison approach, (i) the item-solution and parcel-solution now both have nearly the same power, and (ii) the item-solution and parcel-solution now both have much higher power (~ 0.80) than when using Rhemtulla’s (2016) global χ^2 test of absolute fit for detecting a structural misspecification. These points are demonstrated in Figure 9 for detecting the structural misspecification in the Situation 3 running example. Note that in this running example, theoretical power for the model comparison can be computed using a special case of Satorra and Saris (1985) rather than using the nested model procedure of Satorra and Saris (1983) because Model B is in fact also the generating model. Key features of these results are explained as follows.

- a. The reason that, using the structural model comparison approach, the item-solution and parcel-solution yield virtually equal power (within rounding) in Figure 9 for detecting the structural misspecification is that there is again nearly identical noncentrality for the item- and parcel-level analyses (because it arises from the structural misspecification (noncentrality = 8 for both)) but additionally there is now the same Δdf for both the item- and parcel-analyses ($\Delta df = 1$). This is represented visually in Figure 10.

¹³Regarding the closer correspondence between theoretical and empirical power for detecting the structural misspecification using a χ^2 difference test than using the χ^2 test of absolute fit in Figure 9, it can be shown that the empirical inflation of the χ^2 statistic under low N and high p is smaller for the χ^2 difference test. Note that the empirical power calculation also allows the researcher to quantify and report PAV, as in Figure 9.

- b. The reason that the power for the χ^2 difference test in Figure 9 is much higher than the power for the χ^2 test of absolute fit of the misspecified Model B is that the noncentrality for the χ^2 difference test is equal to the noncentrality of the χ^2 test of absolute fit of Model A in this circumstance (MacCallum, Browne, & Cai, 2006; Steiger, Shapiro, & Browne, 1985)¹⁴ but the Δdf for the difference test (i.e., 1) is much lower than the df under *either* test of absolute fit (i.e., 592 for the item solution and 25 for the parcel-solution) (relatedly, see Mulaik et al., 1989). This is represented visually in Figure 10.

Implications for practice

The desire to use parceling to improve power for detecting structural misspecification when assuming a correct measurement model does not motivate parceling without considering PAV. Specifically, in contrast to Situation 3, for a gain in power for detecting structural misspecifications, it is not necessary to use a parcel-solution plus a global test of absolute fit. Rather, even higher power can be obtained by comparing models differing in structural relations, regardless of whether parcel-solutions or item-solutions are used. Further, using item-solutions for such model comparisons can yield the same power as parcel-solutions, on average across allocations, while avoiding PAV (which is marked under some settings—Sterba & Rights, 2017).

Discussion

Here, we described and critiqued rationales for three situations under which parcel-allocation variability

(PAV) has been ignored in recent methodological literature. None were found to offer a viable reason to ignore PAV. Doing so raises both representativeness and replicability concerns about a parceling study. Our field is grappling with concerns about replicability, recently voiced in the context of parcel-solutions (Maul, 2012). Acknowledging and reporting information about PAV in parcel-solutions can help relate and synthesize results of parcel-solutions across studies.

Summary

In sum, in Situation 1 the existence of PAV was thought to be defined away by restricting focus to a kind of parceling strategy (purposive) with only one possible allocation, thus precluding the creation of a PAV distribution of results across alternative parcel-allocations. We showed that a purposive algorithm can still be repeatedly implemented within-sample, and is still subject to PAV. Ignoring PAV in this context obfuscated comparisons of results between purposive algorithms. In Situation 2 the need to quantify PAV was thought to be bypassed by using an item-level test (for unidimensionality) that would permit parceling when PAV wouldn't be a concern. However, PAV was shown to still arise when such a test is passed; this same general pattern of results would hold for other kinds of tests of unidimensionality as well. Situation 3 sought to identify new benefits of parceling—in terms of improved power for detecting structural misspecification—while ignoring PAV. Applied researchers could have interpreted this new motivation for parceling as counterbalancing the perceived drawback of PAV. However, in empirical parceling practice it was shown that this benefit may not actually arise, though PAV does arise. Furthermore, an alternative testing approach was shown to provide a much greater improvement in power regardless of whether an item-solution is used (thus, avoiding PAV altogether) or a parcel-solution is used.

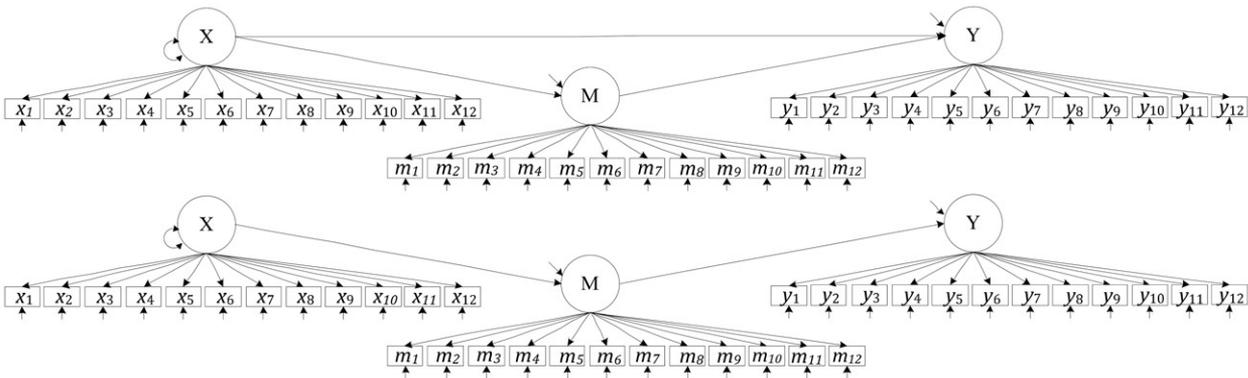
In the following section, we address relationships among the situations by considering a few incompatibilities. Subsequently, we conclude with recommendations for methodological and empirical research involving parceling.

Relationships among the situations

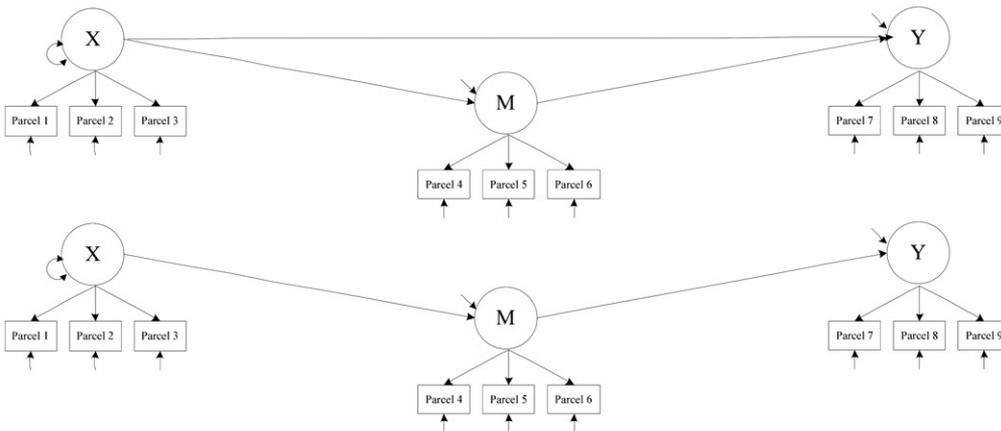
Above we addressed Situations 1–3 separately because they were proposed separately in the literature. We also did so because, in the view of their respective

¹⁴From Steiger et al. (1985) Theorem 1, the chi-square test of absolute fit of Model A uses test statistic $n\hat{F}^{(A)} \sim$ noncentral χ^2 with $df = v_A$ and noncentrality δ_A whereas the chi-square difference test for Model A versus Model B uses test statistic $[(n\hat{F}^{(A)} - n\hat{F}^{(B)})] \sim$ noncentral χ^2 with $df = (v_A - v_B)$ and noncentrality $(\delta_A - \delta_B)$. Here, $\hat{F}^{(A)}$ and $\hat{F}^{(B)}$ are minimized discrepancies for a sample size of n for Models A and B, respectively. Noncentralities δ_B and δ_A are population “badness of fit” quantities. If both Model B and A are incorrect, the noncentralities for these tests are different (i.e., δ_A vs. $(\delta_A - \delta_B)$). However, in the case where Model B is the generating model (as also assumed in Rhemtulla (2016), and in widespread parceling practice when researchers routinely assume no measurement model error and calculate power for a particular parametric structural misspecification), $\delta_B = 0$ and $\delta_A > 0$. In this case, the chi-square test of absolute fit of Model A uses test statistic $n\hat{F}^{(A)}$ with $df = v_A$ and noncentrality δ_A whereas the chi-square difference test for Model A versus B uses test statistic $(n\hat{F}^{(A)} - n\hat{F}^{(B)})$ with $df = (v_A - v_B)$ and noncentrality δ_A . Although the latter two tests use the same noncentrality, that does not imply the same power (see Figures 9 and 10) because although in either test the centers of the null and alternative distributions differ by δ_A , for a given test the pair of (null and alternative) distributions is pushed left or right depending on the values of v_B and v_A (see Figure 10).

Panel I. Fitted (full) item-level Model B and fitted (reduced) item-level Model A used to compute power for detecting the structural misspecification using a $\chi^2(1)$ difference test comparing B and A.



Panel II. Fitted (full) parcel-level Model B and fitted (reduced) parcel-level Model A used to compute power for detecting the structural misspecification using a $\chi^2(1)$ difference test comparing B and A.



Notes. The misspecification was the same as in Figure 7 (an omitted direct effect of the x-factor on the y-factor).

Figure 8. Background on Situation 3: An alternative approach to computing item-level or parcel-level power for detecting a structural misspecification involves comparing models with and without the structural misspecification and assessing power for a test of the difference in fit.

authors, they are not all compatible with each other. As one example, Situation 2 is incompatible with Situation 1 because Situation 1 allows heterogeneous/distributive purposive parceling of multidimensional constructs but Situation 2 does not. According to Marsh et al. (2013, p. 260), the “underlying rationale of the explicit distributive [i.e., heterogeneous] strategy only makes sense when the assumption of unidimensionality is violated, thus, precluding the appropriate use of item parcels.” Our own perspective lies in the middle. That is, multidimensional constructs (such as in Figure 1 Panel A, which often arise in real-world practice) do not themselves preclude parceling, so long as PAV is accounted for and reported.

As another example, Situations 2 and 3 are incompatible with Situation 1 because Situation 1 does not ultimately require a test of the item-level model as a prerequisite for parceling (as do Marsh et al. [2013] and Rhemtulla [2016]), but rather would still allow researchers to rely on substantive justification for the correct specification of the measurement portion of the model prior to parceling. According to Marsh et al. (2013, p. 281), “whereas we focus on empirical tests of when parceling is or is not appropriate, Little et al. provides little in the way of testable criteria to justify the use of parcels.” Again, our perspective lies in the middle. A test of absolute fit of the item-level model can indeed be helpfully informative

		What test?	
		Test of absolute fit	Test of difference in fit
		<ul style="list-style-type: none"> • for model with structural constraints • assuming correct measurement model • used χ^2 test of absolute fit of Model A 	<ul style="list-style-type: none"> • for models with vs. without structural constraints • same measurement model • used χ^2 difference test comparing Models A & B
What method to compute power?	Theoretical Power* (e.g., Satorra & Saris, 1985)	Item-level power: .08 Parcel-level power: .28 (used in Rhemtulla, 2016)	Item-level power: .81 Parcel-level power: .81
	Empirical Power*	Item-level power: .63 Parcel-level power: .29 PAV: range=.27-.33	Item-level power: .79 Parcel-level power: .79 PAV: range=.77-.80 (used in Sterba & Rights, 2017)

Notes. PAV = parcel allocation variability. Models A and B were defined in Figure 8. The misspecification we are trying to detect is the same as in Rhemtulla (2016) (i.e., omitting the direct effect of the x-factor on the y-factor in Figure 6). * The theoretical and empirical power analysis procedures were described in the manuscript.

Figure 9. Evaluation of Situation 3: Power for detecting the same structural misspecification at the same *N* using two different levels of analysis (item vs. parcel), two different testing approaches (absolute fit vs. difference in fit), and two different methods of computing power (theoretical vs. empirical).

about the location and extent of measurement model error. However, if applied researchers are considering parceling specifically because the item-level model is inestimable (which occurs more often with complex models and at lower sample sizes), we would not bar them from parceling, as would Situation 2. Applied researchers have limited analytic options available to them when their item-level model is inestimable, and parceling could be allowable so long as PAV were investigated and accounted for. Measurement model error in a parcel-level analysis that remains undiagnosed due to an inestimable item-level model will to some extent contribute to and be reflected in PAV.

Recommendations for methodological research on parceling

The demonstrations presented throughout this manuscript lead to the following two recommendations for future methodological studies on parceling.

1. Methodological research and simulation studies on parceling should not avoid sampling error by focusing exclusively on the performance of parceling in the population. This circumvents consideration of PAV in an unrealistic manner which does not mirror empirical parceling practice. In practice, researchers parcel using sample-level data, not census data, and moreover they parcel

much more often when they have lower communalities and/or lower sample size, which in turn increases the magnitude of PAV, all else equal (Sterba, 2011; Sterba & MacCallum, 2010). These realistic conditions should be included in parceling simulation designs, in conjunction with evaluating PAV. Otherwise, simulations provide researchers with results and recommendations for using parceling that do not reflect the complexities of parceling in empirical practice, as was demonstrated in earlier sections.

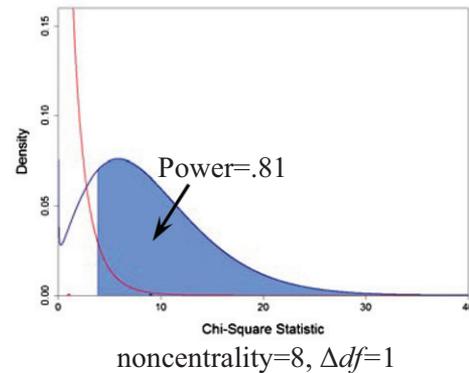
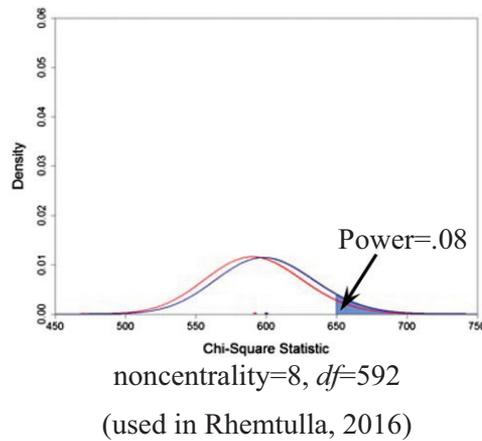
2. Methodological research and simulation studies on parceling should not ignore PAV because of the fact that they involve purposive parceling, multidimensional constructs, and/or no measurement error. PAV can arise in empirical practice under all of these settings. Hence, it is again unrealistic to provide researchers with results and recommendations for using parceling that do not factor in the uncertainty in parcel-solution results due to PAV.

Addressing these recommendations in methodological research requires modifying common simulation designs. Commonly, Monte Carlo simulation study designs allow assessing bias and variability of parcel-solution results across repeated samples using a single parcel-allocation to represent a given parceling algorithm. Investigating PAV requires repeatedly generating parcel-allocations from that parceling

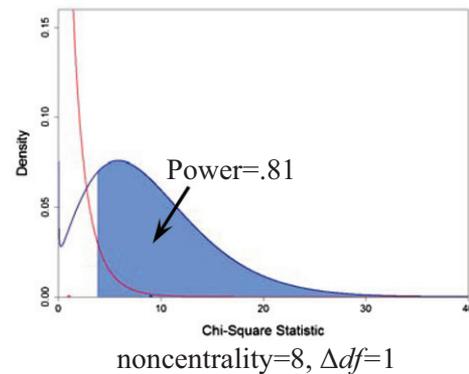
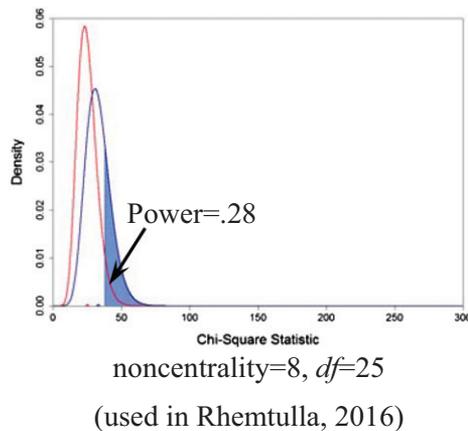
Power for a global χ^2 test of absolute fit of the reduced model (A) that has structural constraint

Power for a χ^2 difference test comparing reduced model A & full model B that differ in structural constraints

Item-level analysis:



Parcel-level analysis:



Notes. The scaling of the axes is the same for plots in a given column. In each plot, the density on the left is the chi-square distribution when the null is true and the density on the right is the chi-square distribution when the null is false to a specified degree (here, corresponding to the misspecification of omitting the direct effect of the x-factor on the y-factor in Figure 6). The null distribution is centered over df . The alternative distribution is centered over $df + \text{noncentrality}$. Even though noncentrality is the same in both plots in column 1, the distributions in the top and bottom panels look different because of the shape of the chi-square distribution at those different df s. In contrast, in column 2, the noncentrality and df are the same in both plots, so the plots in column 2 are identical.

Figure 10. Visual representation of the theoretical power comparison reported in Figure 9 for detecting the same structural misspecification at the same N .

algorithm within each of these samples, and then fitting the model of interest to each parcel-allocation within each sample. This procedure allows quantifying and distinguishing the magnitude and effects of sampling variability versus parcel-allocation variability (for details see Sterba, 2011; Sterba & MacCallum, 2010; or Sterba & Rights, 2017).

Recommendations for empirical applications using parceling

In applied practice, initial reactions to the concept of PAV ranged from considering it a nuisance and a hassle to investigate, to reacting with concern and avoidance—concern that merely mentioning the possibility

or existence of PAV could spook reviewers and threaten publication, leading to avoidance of addressing PAV. We have been told about researchers searching for “phrasing that doesn’t invite any suggestions to reallocate parcels and re-run” to avoid the situation where “a reviewer may ask you to try out a few other allocation schemes and determine how consistent the results are. That’s a big can of worms.”

Historically, when confronted with new sources of uncertainty in statistical results, researchers have likewise reacted with concern that was followed by acceptance only after widespread pedagogical dissemination informed reviewers and journal editors about how the uncertainty could be quantified and addressed (Panter & Sterba, 2011). One historical example concerns missing data uncertainty. When faced with missing data uncertainty, researchers were initially reluctant to repeatedly generate random multiple imputations and pool results across imputations, as results could differ when a different set of imputations were employed. Rubin (1996, p. 479) explains that “an early criticism, not much heard anymore but worthy of response, is that multiple imputation is theoretically unsatisfactory and practically unacceptable because it adds random noise to the data. In this context, it is critical to remember that multiple imputation does not pretend to *create* information through simulated values but simply to *represent* the observed information this way so as to make it amenable to valid analysis... The extra noise created when using a finite number of imputations is the price to be paid for this luxury.” Other early concerns about multiple imputation according to Rubin (1996, p. 480) were that the “multiply imputed data sets take too much storage” and “multiple imputation [is] too much work for the user.” Software advances and accessible pedagogical treatments of multiple imputation (e.g., Enders, 2010) have allayed these concerns. Now ignoring missing data uncertainty in applied research would be widely considered unacceptable by reviewers and journal editors.

Our recommendations for applied practice continue to involve, first, acknowledging the possibility of PAV and, second, accounting for and reporting PAV. These recommendations have been detailed elsewhere, but we provide a brief overview here. One approach to implement these recommendations involves conducting a sensitivity analysis (see Sterba & MacCallum, 2010; Sterba & Rights, 2017 for procedures). For instance, researchers can investigate and report whether substantive conclusions based on inferential decisions about parcel-solution overall fit,

model ranking, or individual/multiparameter tests would change across many repeated item-to-parcel allocations (e.g., 500) within their sample and can examine the magnitude of PAV in point estimates of fit indices and structural model parameters. If overall substantive conclusions do not change across alternative parcel allocations, then they are robust to the existence of PAV. If overall substantive conclusions can change across alternative parcel allocations, this degree of uncertainty can be acknowledged, and future studies can seek to minimize it by reducing sampling error and/or model error.

Another approach to implementing these recommendations involves pooling results across allocations (see Sterba & Rights, 2016 for procedures). This approach can be used, for instance, to provide a single (pooled) point estimate for each structural parameter (rather than a range of point estimates across repeated allocations within sample). This approach can also be used to provide a single (pooled) standard error for each structural parameter that combines sources of uncertainty stemming from both sampling variability and parcel-allocation variability. Thus, this approach yields a single inferential decision for each structural parameter, rather than a range of (potentially significant to nonsignificant) results across repeated allocations within sample. As a concrete example, suppose an applied researcher were interested in an inferential decision about the indirect effect (a-path \times b-path) when fitting the Figure 6 model (where the a-path is the *x*-factor’s effect on the *m*-factor and the b-path is the *m*-factor’s effect on the *y*-factor). This researcher has a single sample, which we drew from the Situation 3 simulation. In this sample, the indirect effect is significant in 66% of allocations and nonsignificant in 34% of allocations; that is, there is PAV in the substantively important inferential decision about the indirect effect.¹⁵ However, using Sterba and Rights’ (2016) pooling approach we can obtain a single inferential decision that accounts for both sampling variability and PAV: a 95% CI of {-.002–.242}, indicating that the null hypothesis is rejected. Here, the pooled a-path SE, pooled a-path estimate, pooled b-path SE, and pooled b-path estimate were used, together with conventional procedures in Preacher and Selig (2012), for obtaining a Monte Carlo CI for an indirect effect. Note that Sterba and Rights (2016)

¹⁵For reference purposes, note that 40% of samples from the Situation 3 simulation similarly had PAV in the result of the significance test for the indirect effect. Significance of the indirect effect was determined by creating a 95% Monte Carlo confidence interval for the indirect effect (using Preacher & Selig, 2012 procedures), for each allocation, and then checking to see if that 95% CI includes the null hypothesized value of 0.

also supplied supplementary diagnostics, termed PPAV and RPAV, for quantifying the degree of uncertainty in estimates that is due to parcel allocation variability; PPAV indicates the proportion of total variance of a parameter estimate that is attributable to parcel-allocation variability and RPAV indicates the ratio of parcel-allocation variability to sampling variability in a parameter estimate.

Investigations of PAV are beginning to be employed and reported in empirical parceling applications (e.g., Ayturk, 2016; Cole et al., 2017; Cole et al., 2018; Kam & Meyer, 2015; Sainio et al., 2013; Trautwein et al., 2015). As one concrete example, in Cole et al. (2017), 20% of the variability in structural model coefficients was found to be attributable to PAV, leading to some different inferences about structural model parameters when using repeated single-allocations, as compared to a pooled-allocation approach that accounted for PAV and yielded a single inferential decision. More generally, across empirical studies to date, the impact and magnitude of PAV has ranged from a “small” amount of PAV (e.g., Kam & Meyer, 2015),¹⁶ to results being “remarkably influenced” by PAV (Ayturk, 2016). For the widely held goals of replicability and representativeness of results, it is beneficial to know such information about PAV for studies using parceling.

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.

¹⁶Kam and Meyer (2015) reported the within-sample across-allocations average of the RMSEA to be .06, but did not report the percentage of allocations in which the test of close fit was rejected. Note that even if the standard deviation of the PAV distribution appears small, when the PAV distribution is centered so near to the decision threshold of close versus not close fit, there can be practically meaningful changes in results across allocations in the form of fit flipping between close and not close across allocations (as was shown in Figure 5).

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: This article is based on a Cattell Award address given to the Society of Multivariate Experimental Psychology in October, 2016. The authors would like to thank Kristopher J. Preacher, Jason D. Rights, and Robert C. MacCallum for helpful comments on this work. Correspondence concerning this article should be addressed to Sonya Sterba, Department of Psychology and Human Development, Vanderbilt University, Peabody #552, 230 Appleton Place, Nashville, TN 37203. Email: Sonya.Sterba@Vanderbilt.edu. The ideas and opinions expressed herein are those of the author alone, and endorsement by Vanderbilt University is not intended and should not be inferred.

References

- Asparouhov, T., & Muthén, B. (2010). *Plausible values for latent variables in Mplus*. Technical document Retrieved from www.statmodel.com.
- Ayturk, E. (2016). *The product indicator approach to estimation of latent interaction effects: Testing of a new method* (Unpublished masters thesis). Fordham University, New York.
- Bagozzi, R. P., & Edwards, J. R. (1998). A general approach for representing constructs in organizational research. *Organizational Research Methods, 1*(1), 45–87. doi:10.1177/109442819800100104
- Bagozzi, R. P., & Heatherton, T. F. (1994). A general approach to representing multifaceted personality constructs: Application to state self-esteem. *Structural Equation Modeling, 1*(1), 35–67. doi:10.1080/10705519409539961
- Bandalos, D. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling, 9*(1), 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*(2), 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. doi:10.4324/9781410601858
- Bonifay, W., Reise, S., Scheines, R., & Meijer, R. (2015). When are multidimensional data unidimensional enough for structural equation modeling? An evaluation of the DETECT multidimensionality index. *Structural Equation*

- Modeling*, 22(4), 504–516. doi:10.1080/10705511.2014.938596
- Braun, H., & von Davier, M. (2017). The use of test scores from large-scale assessment surveys: Psychometric and statistical considerations. *Large-Scale Assessments in Education*, 5, 5–17. doi:10.1186/s40536-017-0050-x
- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.
- Cattell, R. B. (1956). A shortened “Basic English” version (Form C) of the 16 PF Questionnaire. *The Journal of Social Psychology*, 44(2), 257–278. doi:10.1080/00224545.1956.9921928
- Coffman, D. L., & MacCallum, R. C. (2005). Using parcels to convert path analysis models into latent variable models. *Multivariate Behavioral Research*, 40(2), 235–259. doi:10.1207/s15327906mbr4002_4
- Cole, D. A., Perkins, C. E., & Zerkowicz, R. L. (2016). Impact of homogeneous and heterogeneous parceling strategies when latent variables represent multidimensional constructs. *Psychological Methods*, 21(2), 164–174. doi:10.1037/met0000047
- Cole, D. A., Martin, J. M., Jacquez, F. M., Tram, J. M., Zerkowicz, R., Nick, E. A., & Rights, J. D. (2017). Time-varying and time-invariant dimensions of depression in children and adolescents: Implications for cross-informant agreement. *Journal of Abnormal Psychology*, 126(5), 635–651. doi:10.1037/abn0000267
- Cole, D. A., Goodman, H. J., Garber, J., Cullum, K. A., Cho, S.-J., Rights, J. D., ... Simon, H. F. M. (2018). Validating parent and child forms of the Parent Perception Inventory. *Psychological Assessment*, 30, 1065–81. doi:10.1037/pas0000552
- Costa, P. T., & McCrae, R. R. (1985). *The NEO personality inventory: Manual (Form S and Form R)*. Odessa, FL: Psychological Assessment Resources, Inc.
- Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford.
- Finch, H., & Habing, B. (2007). Performance of DIMTEST and NOHARM based statistics for testing unidimensionality. *Applied Psychological Measurement*, 31(4), 292–307. doi:10.1177/0146621606294490
- Gerbing, D., & Anderson, J. (1988). An updated paradigm for scale development incorporating unidimensionality and its assessment. *Journal of Marketing Research*, 25(2), 186–192. doi:10.2307/3172650
- Gribbons, B. C., & Hocevar, D. (1998). Levels of aggregation in higher level confirmatory factor analysis: Application for academic self-concept. *Structural Equation Modeling*, 5(4), 377–390. doi:10.1080/10705519809540113
- 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(2), 168–193. doi: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(3), 233–256. doi:10.1177/109442819923002
- Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9(2), 139–164. doi:10.1177/014662168500900204
- 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(2), 327–351. doi:10.1111/j.2044-8317.2004.tb00142.x
- Hoskens, M., & De Boeck, P. (1997). A parametric model for local dependence among test items. *Psychological Methods*, 2(3), 261–277. doi:10.1037//1082-989x.2.3.261
- Irwing, P., Booth, T., & Batey, M. (2014). An investigation of the factor structure of the 16PF, Version 5: A confirmatory factor and invariance analysis. *Journal of Individual Differences*, 35(1), 38–46. doi:10.1027/1614-0001/a000125
- Kam, C., & Meyer, J. (2015). Implications of item keying and item valence for the investigation of construct dimensionality. *Multivariate Behavioral Research*, 50(4), 457–469. doi:10.1080/00273171.2015.1022640
- Kim, S., & Hagtvet, K. A. (2003). The impact of misspecified item parceling on representing latent variables in covariance structure modeling: A simulation study. *Structural Equation Modeling*, 10(1), 101–127. doi:10.1207/s15328007sem1001_5
- Kishton, J. M., & Widaman, K. F. (1994). Unidimensional versus domain representative parceling of questionnaire items: An empirical example. *Educational and Psychological Measurement*, 54(3), 757–765. doi:10.1177/0013164494054003022
- Landis, R. S., Beal, D. J., & Tesluk, P. E. (2000). A comparison of approaches to forming composite measures in structural equation modeling. *Organizational Research Methods*, 3(2), 186–207. doi:10.1177/109442810032003
- Latane, B. (1989). Social psychology and how to revitalize it. In M. R. Leary (Ed.), *The state of social psychology: Issues, themes, and controversies* (pp. 1–12). Newbury Park, CA: Sage.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley. doi:10.1002/9781119013563
- 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(2), 151–173. doi: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(3), 285–300. doi:10.1037/a0033266
- MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological Methods*, 11(1), 19–35. doi:10.1037/1082-989x.11.1.19
- 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(3), 257–284. doi:10.1037/a0032773
- Martens, M. (2005). The use of structural equation modeling in counseling psychology research. *The Counseling Psychologist*, 33(3), 269–298. doi:10.1177/0011000004272260
- Matsunaga, M. (2008). Item parceling in structural equation modeling: A primer. *Communication Methods and Measures*, 2(4), 260–293. doi:10.1080/19312450802458935

- Maul, A. (2012). Examining the structure of emotional intelligence at the item level: New perspectives, new conclusions. *Cognition and Emotion*, 26(3), 503–520. doi:10.1080/02699931.2011.588690
- Meade, A. W., & Kroustalis, C. M. (2006). Problems with item parceling for confirmatory factor analytic tests of measurement invariance. *Organizational Research Methods*, 9(3), 369–403. doi:10.1177/1094428105283384
- Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56(2), 177–196. doi:10.1007/bf02294457
- Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin*, 105(3), 430–445. doi:10.1037/0033-2909.105.3.430
- Nasser, F., & Wisenbaker, J. (2003). A Monte Carlo study investigating the impact of item parceling on measures of fit in confirmatory factor analysis. *Educational and Psychological Measurement*, 63, 729–57.
- 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(2), 204–228. doi:10.1207/s15328007sem1302_3
- Panter, A. T., & Sterba, S. K. (Eds.). (2011). *Handbook of ethics in quantitative methodology*. New York: Taylor & Francis/Routledge. doi:10.4324/9780203840023
- Plotly Technologies Inc. (2015). *Collaborative data science*. Montréal, QC: Plotly Technologies Inc.
- 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.
- Preacher, K. J., & Selig, J. P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures*, 6(2), 77–98. doi:10.1080/19312458.2012.679848
- Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement*, 73(1), 5–26. doi:10.1177/0013164412449831
- Reiter, J., & Raghunathan, T. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102(480), 1462–1471. doi:10.1198/016214507000000932
- Rhemtulla, M. (2016). Population performance of SEM parceling strategies under measurement and structural model misspecification. *Psychological Methods*, 21(3), 348–368. doi:10.1037/met0000072
- 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(3), 379–412. doi:10.1207/s15327906mbr3903_1
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley. doi:10.1002/9780470316696
- Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434), 473–489. doi:10.2307/2291635
- Sainio, M., Veenstra, R., Little, T. D., Karna, A., Rönkkö, M., & Salmivalli, C. (2013). Being bullied by same- versus other-sex peers: Does it matter for adolescent victims? *Journal of Clinical Child & Adolescent Psychology*, 42(4), 454–456. doi:10.1080/15374416.2013.769172
- Sass, D., & Smith, P. (2006). The effects of parceling unidimensional scales on structural parameter estimates in structural equation modeling. *Structural Equation Modeling*, 13, 566–86.
- Satorra, A., & Saris, W. E. (1983). The accuracy of a procedure for calculating the power of the likelihood ratio test as used within the LISREL framework. In C. P. Middelndorp, B. Niemoller, & W. E. Saris (Eds.), *Sociometric research 1982* (pp. 127–190). Amsterdam: Sociometric Research Foundation.
- Satorra, A., & Saris, W. E. (1985). The power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, 50(1), 83–90. doi:10.1007/bf02294150
- Schallow, J. (2000). A comparison of three approaches to constructing item parcels to improve subject-to-parameter ratios in confirmatory factor analysis. *Multivariate Experimental Clinical Research*, 12, 29–41.
- Schofield, L. S., Junker, B., Taylor, L. J., & Black, D. A. (2015). Predictive inference using latent variables with covariates. *Psychometrika*, 80(3), 727–747. doi:10.1007/s11336-014-9415-z
- Shi, D., Lee, T., & Terry, R. A. (2018). Revisiting the model size effect in structural equation modeling. *Structural Equation Modeling*, 25(1), 21–40. doi:10.1080/10705511.2017.1369088
- Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika*, 50(3), 253–264. doi:10.1007/bf02294104
- Sterba, S. K. (2011). Implications of parcel-allocation variability for comparing fit of item-solutions and parcel-solutions. *Structural Equation Modeling*, 18(4), 554–577. doi: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(2), 322–358. doi:10.1080/00273171003680302
- Sterba, S. K., & Rights, J. D. (2016). Accounting for parcel-allocation variability in practice: Combining sources of uncertainty and choosing the number of allocations. *Multivariate Behavioral Research*, 51(2–3), 296–313. doi:10.1080/00273171.2016.1144502
- Sterba, S. K., & Rights, J. D. (2017). Effects of parceling on model selection: Parcel-allocation variability in model ranking. *Psychological Methods*, 22(1), 47–68. doi:10.1037/met0000067
- Stucky, B. D., Gottfredson, N. C., & Panter, A. T. (2012). Item-level factor analysis. In H. Cooper, P. Camic, D. Long, A. T. Panter, D. Rindskopf, & K. Sher (Eds.), *APA handbook of research methods in psychology* (Vol. 1, pp. 683–697). Washington, DC: APA Books. doi:10.1037/13619-000
- Takahashi, T., & Nasser, F. (1996, April). *The impact of using item parcels on ad hoc goodness of fit indices in confirmatory factor analysis: An empirical example*. Paper presented at the annual meeting of

- the American Educational Research Association, New York, NY.
- Thompson, B., & Melancon, J. G. (1996, November). *Using item 'testlets'/'parcels' in confirmatory factor analysis: An example using the PPSDQ-78*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Tuscaloosa, AL.
- Trautwein, U., Lüdtke, O., Nagy, N., Lenski, A., Niggli, A., & Schnyder, I. (2015). Using individual interest and conscientiousness to predict academic effort: Additive, synergistic, or compensatory effects? *Journal of Personality and Social Psychology*, *109*(1), 142–162. doi:10.1037/pspp0000034
- van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL: Chapman & Hall/CRC Press.
- 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.
- 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*(4), 233–242. doi:10.1016/j.hrmr.2008.07.002
- Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*(2–3), 114–128. doi:10.1016/j.stueduc.2005.05.005
- 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. doi:10.1177/0146621609338592
- Yuan, K.-H., Tian, Y., & Yanagihara, H. (2015). Empirical correction to the likelihood ratio statistic for structural equation modeling with many variables. *Psychometrika*, *80*(2), 379–405. doi:10.1007/s11336-013-9386-5
- Yuan, K.-H., Zhang, Z., & Zhao, Y. (2017). Reliable and more powerful methods for power analysis in structural equation modeling. *Structural Equation Modeling*, *24*(3), 315–330. doi:10.1080/10705511.2016.1276836