Structural Equation Modeling, 18:554–577, 2011

ISSN: 1070-5511 print/1532-8007 online
DOI: 10.1080/10705511.2011.607073

Psychology Press
Taylor & Francis Group

Implications of Parcel-Allocation Variability for Comparing Fit of Item-Solutions and Parcel-Solutions

Sonya K. Sterba
Department of Psychology and Human Development, Vanderbilt University

This article relates a still-popular motivation for using parceling to an unrecognized cost. The still-popular motivation is improvement in fit with respect to the item-solution. The cost is uncertainty in fit due to the selection of one out of many possible item-to-parcel allocations. A theoretical framework establishes the reason for this relationship: The same mechanisms that cause larger item- versus parcel-solution differences in the minimized discrepancy function also cause larger allocation to allocation variability in the parcel-solution's minimized discrepancy function. Study 1 illustrates that these shared causal mechanisms lead to a strong positive association between average item–parcel differences in minimized discrepancy function values and parcel-allocation variability in those values. Study 2 extends these results from discrepancy function values to fit indexes, showing that the association remains positive, but varies in magnitude depending on what quantities other than the discrepancy function are involved in computing the fit index. The important implication for practice is that when item–parcel fit differences are large enough to alter conclusions about model adequacy, parcel-allocation variability tends to be large enough for parcel-solution model adequacy to depend on the particular allocation chosen.

Keywords: parceling, parcel-allocation variability, model fit, confirmatory factor analysis

Parceling involves the averaging or summing of several raw items to form a single score, which can then be used as an indicator of a latent variable in a factor analysis model or structural equation model. Bandalos and Finney's (2001) and Williams and O'Boyle's (2008) surveys of the use of parceling in applied social science research indicate that parceling is widely used, sometimes with the explicitly stated motivation of improving fit over the item-solution. Indeed, Williams and O'Boyle (2008) explained that

Correspondence should be addressed to Sonya K. Sterba, Department of Psychology and Human Development, Vanderbilt University, Peabody #552, 230 Appleton Place, Nashville, TN 37203, USA. E-mail: Sonya.Sterba@Vanderbilt.edu

> In terms of improved model fit, the advantages of parcels are also important. One of the biggest challenges HRM [Human Resource Management] researchers face in getting their SEM papers published is demonstrating that their model adequately fits the data. A general step in convincing reviewers of a properly specified model involves the use of various fit indices and demonstrating that the model meets the "gold standards" associated with the indices (for example, the recommended .05–.08 range for the RMSEA). . . . So, given this, one can understand the desire of HRM researchers to use parcels. (p. 240)

Bandalos and Finney's (2001) survey found examples of this stated motivation in Bagozzi and Edwards (1998), Thompson and Melancon (1996), and Takahashi and Nasser (1996); other examples are Landis, Beal, and Tesluk (2000), Martens (2005), Rogers and Schmitt (2004), and Nasser and Wisenbaker (2003).

Several reasons for improved fit of parcel-solutions as compared to item-solutions are often reported (e.g., Bandalos, 2002), and will be reviewed later in detail. Reliance on expected improvement in fit as a motivation for employing parceling has been discouraged because such improvements can mask, or occur in spite of, measurement model misspecifications (Bandalos, 2002; Bandalos & Finney, 2001; Hall, Snell, & Foust, 1999; Lee, 2005). This critique led Bandalos and Finney (2001) to state that "the crucial factor in a researcher's decision to use item parceling appears to be the degree to which he or she is willing to make the assumption that the use of item parceling has not masked any substantively and/or theoretically important sources of lack of fit" (p. 286). This article shows that even if a researcher *correctly* makes the assumption that there is no measurement model error, choosing parcel-solutions based on changes in fit vis-à-vis item solutions is problematic based on previously unrecognized grounds.

In this article, I first review theoretical background for changes in fit between item-solutions and parcel-solutions. Second, drawing on this theoretical framework, I hypothesize that the typical *difference* in the minimized discrepancy function value between the item-solution and the parcel-solution should be positively and monotonically related to the *variability* in that minimized discrepancy function across alternative parcel-solutions (called *parcel-allocation variability;* Sterba & MacCallum, 2010[1] and described later). I illustrate this point using a simulation. Third, I show that item–parcel differences in model fit indexes are also positively related to the amount of parcel-allocation variability in that fit index. But the particular nature and magnitude of their relation depends on certain factors, such as what information in addition to the discrepancy function value the fit index contains (e.g., sample size or penalties for number of parameters). A major implication of these findings for applied practice is that researchers should expect larger observed changes in fit between item-solutions and parcel-solutions to be accompanied, on average, by interjection of unwanted uncertainty into the parcel-solution estimates in the form of allocation-to-allocation variability in model fit statistics. This finding introduces an additional reason why, even for a properly specified model, it is an ill-conceived practice to choose parcel-solutions over item-solutions based on changes in fit—particularly without taking into account the accompanying parcel-allocation variability (discussed later).

[1] Sterba and MacCallum's (2010) paper was solely focused on introducing the concept of parcel-allocation variability and documenting its occurrence in the context of alternative parcel-solutions. Their simulation did not consider item-solutions, nor did they relate the concept of parcel-allocation variability to differences between item- and parcel-solutions, as done here.

REVIEW OF THEORETICAL FRAMEWORK

Recall that during model estimation for factor analysis models and SEMs, a function is minimized that quantifies the discrepancy between observed covariances and model-implied covariances.[2] When evaluated at the final parameter estimates, this *discrepancy function* is at its lowest value.[3] We denote $\hat{F}_p$ as the minimized discrepancy function value for a parcel-solution (where discrepancy between observed and model-implied parcel covariances is minimized) and $\hat{F}_i$ as the minimized discrepancy function value for an item-solution (where discrepancy between the observed and model-implied *item* covariances is minimized).

The reasons for differences between $\hat{F}_i$ and $\hat{F}_p$ are direct consequences of the theoretical framework of MacCallum and Tucker (1991), as has been previously stated by Bandalos (2002), Hall et al. (1999), and Sass and Smith (2006).[4] However, using the extension of this framework to include parcel-solutions, from Sterba and MacCallum (2010), we are able to relate the typical difference between $\hat{F}_i$ and $\hat{F}_p$, on the one hand, to the variation in $\hat{F}_p$ across alternative allocations of items to parcels, on the other hand. I begin with a brief review of the original framework and its extension.[5]

An item-level factor analytic population covariance structure for a measurement model is

$$E(x_i x_i') = \boldsymbol{\Sigma}_i = \boldsymbol{\Lambda}_i \boldsymbol{\Phi}_i \boldsymbol{\Lambda}_i' + \boldsymbol{\Psi}_i^2 \tag{1}$$

where i denotes *item-level*, E is the expectation operator, x_i is an $m \times 1$ vector of mean-centered items in the population, $\boldsymbol{\Lambda}_i$ is an $m \times q$ common factor loading matrix, $\boldsymbol{\Phi}_i$ is a $q \times q$ covariance matrix of common factors, and $\boldsymbol{\Psi}_i^2$ is a (usually) diagonal $m \times m$ matrix of unique variances. In the population, covariances of common and unique factors and covariances of unique factors are assumed to be 0. However, in a given sample of items generated by Equation 1, covariances among common and unique factors—denoted $\mathbf{C}_{uc_i}$ and $\mathbf{C}_{cu_i}$—and covariances among unique factors—denoted $\mathbf{C}_{uu_i}$—will usually not be zero, due to sampling variability. We thus can represent these covariances explicitly in the item-level *sample* covariance structure, $\mathbf{C}_i$:

$$\mathbf{C}_i = \boldsymbol{\Lambda}_i \mathbf{C}_{cc_i} \boldsymbol{\Lambda}_i' + \boldsymbol{\Lambda}_i \mathbf{C}_{cu_i} \boldsymbol{\Psi}_i' + \boldsymbol{\Psi}_i \mathbf{C}_{uc_i} \boldsymbol{\Lambda}_i' + \boldsymbol{\Psi}_i \mathbf{C}_{uu_i} \boldsymbol{\Psi}_i' \tag{2}$$

where $\mathbf{C}_{cc_i}$ is the sample covariance matrix of common factors (MacCallum & Tucker, 1991).

A parcel-level factor analytic population covariance structure is constructed in the following manner. We first form an $k \times m$ selection matrix $\mathbf{A}$ that selects m items for k parcels (see Sterba & MacCallum, 2010, for a worked example). Then $x_p = \mathbf{A}x_i$ is a vector of parcels, of order k. Given a fixed, prespecified number of parcels per factor and items per parcel, locations of nonzero elements in $\mathbf{A}$ determine a particular *item-to-parcel allocation* and can be chosen

[2]The points raised in this article should apply to situations where observed and model-implied means are also included in the discrepancy function, but this situation is not specifically considered here.

[3]One popular discrepancy function is the maximum likelihood discrepancy function. I use it in the simulation described later, but this framework does not apply to just one particular discrepancy function.

[4]These authors used terms that were proxies or related to discrepancy function values in their articles, but it is necessary for the purposes of this presentation to rephrase their comments in terms of discrepancy function values.

[5]This framework is presented in the context of confirmatory factor analysis (CFA), but concepts are generalizable to structural equation models more generally.

by any *random* or *purposive* allocation method available. An example of a purposive item-to-parcel allocation of $m = 6$ items into $k = 3$ two-item parcels is to parcel the pair of items with the highest correlations together, and then parcel the next highest correlated pair, and so on. An example of a random item-to-parcel allocation of the same $m = 6$ items into $k = 3$ two-item parcels is to arbitrarily choose which items to put into each of the parcels—perhaps parceling 1 with 4, 2 with 3, and 5 with 6; or perhaps parceling 3 with 6, 5 with 1, and 2 with 4. Pre- and postmultiplying Equation 1 by $\mathbf{A}$ (i.e., allocating items to parcels) yields the parcel-level *population* covariance structure:

$$E(\mathbf{A}x_i x_i'\mathbf{A}') = \mathbf{A}\boldsymbol{\Sigma}_i\mathbf{A}' = \mathbf{A}\boldsymbol{\Lambda}_i\boldsymbol{\Phi}_i\boldsymbol{\Lambda}_i'\mathbf{A}' + \mathbf{A}\boldsymbol{\Psi}_i^2\mathbf{A}'. \tag{3}$$

Rewriting $x_p = \mathbf{A}x_i$ and $\boldsymbol{\Sigma}_p = \mathbf{A}\boldsymbol{\Sigma}_i\mathbf{A}'$ and $\boldsymbol{\Lambda}_p = \mathbf{A}\boldsymbol{\Lambda}_i$ and $\boldsymbol{\Psi}_p^2 = \mathbf{A}\boldsymbol{\Psi}_i^2\mathbf{A}'$, where a p subscript denotes *parcel-level,* leaves

$$E(x_p x_p') = \boldsymbol{\Sigma}_p = \boldsymbol{\Lambda}_p\boldsymbol{\Phi}_i\boldsymbol{\Lambda}_p' + \boldsymbol{\Psi}_p^2. \tag{4}$$

Including sources of sampling error yields the parcel-level *sample* covariance structure:

$$\mathbf{A}\mathbf{C}_i\mathbf{A}' = \mathbf{A}(\boldsymbol{\Lambda}_i\mathbf{C}_{cc_i}\boldsymbol{\Lambda}_i')\mathbf{A}' + \mathbf{A}(\boldsymbol{\Lambda}_i\mathbf{C}_{cu_i}\boldsymbol{\Psi}_i')\mathbf{A}' + \mathbf{A}(\boldsymbol{\Psi}_i\mathbf{C}_{uc_i}\boldsymbol{\Lambda}_i')\mathbf{A}' + \mathbf{A}(\boldsymbol{\Psi}_i\mathbf{C}_{uu_i}\boldsymbol{\Psi}_i')\mathbf{A}' \tag{5}$$

Finally, rewriting $\mathbf{C}_p = \mathbf{A}\mathbf{C}_i\mathbf{A}'$ and $\boldsymbol{\Lambda}_p = \mathbf{A}\boldsymbol{\Lambda}_i$ and $\boldsymbol{\Psi}_p = \boldsymbol{\Psi}_i'\mathbf{A}' = \mathbf{A}\boldsymbol{\Psi}_i$ we obtain:

$$\mathbf{C}_p = \boldsymbol{\Lambda}_p\mathbf{C}_{cc_i}\boldsymbol{\Lambda}_p' + \boldsymbol{\Lambda}_p\mathbf{C}_{cu_i}\boldsymbol{\Psi}_p' + \boldsymbol{\Psi}_p\mathbf{C}_{uc_i}\boldsymbol{\Lambda}_p' + \boldsymbol{\Psi}_p\mathbf{C}_{uu_i}\boldsymbol{\Psi}_p'. \tag{6}$$

STUDY 1: RELATIONSHIP OF AVERAGE ($\hat{F}_i - \hat{F}_p$) TO PARCEL-ALLOCATION VARIABILITY IN $\hat{F}_p$

In light of this theoretical framework, there can be seen to be several mechanisms through which parcel-solutions can produce smaller $\hat{F}_p$ than item-solutions' $\hat{F}_i$ (e.g., Bandalos, 2002). These mechanisms are reviewed here, by comparing Equations 2 and 6.

- *Mechanism 1.* The first mechanism is the decrease in the number of indicators per factor—which in turn decreases the dimensionality of the $\mathbf{C}_{uu_i}$, $\mathbf{C}_{cu_i}$, and $\mathbf{C}_{uc_i}$ matrices—which then reduces the contribution of error due to unmodeled associations in these matrices. For example, if there are unmodeled error covariances among items in the same parcel (resulting either from sampling error or from measurement model error), parceling decreases their effects on $\hat{F}$ by repackaging their contribution to $\mathbf{C}_{uu_i}$ into common factor variance in $\mathbf{C}_{cc_i}$.
- *Mechanism 2.* The second mechanism through which $\hat{F}_p$ can improve on $\hat{F}_i$ is by reducing the element size and dimensionality of the weight matrix $\boldsymbol{\Psi}_i$, which further decreases the contribution of $\mathbf{C}_{uu_i}$, $\mathbf{C}_{cu_i}$, and $\mathbf{C}_{uc_i}$. The element size reduction occurs because, for instance when item averaging is used to form parcels, the diagonal element of $\boldsymbol{\Psi}_p^2$ for a given parcel is equal to the average of its constituent items' diagonal elements of $\boldsymbol{\Psi}_i^2$, divided by the number of items per parcel.

- *Mechanism 3.* The third mechanism is the reduction in the effects of incorrect constraints (e.g., omitted cross-loadings) in $\mathbf{\Lambda}_i$ on $\hat{F}$ by reapportioning some of their influences into common factor covariances in $\mathbf{C}_{cc_i}$ (e.g., Bandalos, 2002; Hall et al., 1999).

Mechanisms 2 and 3 are not only responsible for the magnitude of $(\hat{F}_i - \hat{F}_p)$ differences, however. Rather, drawing on Sterba and MacCallum (2010), Mechanisms 2 and 3 also can be seen to cause variation in $\hat{F}_p$ across alternate item-to-parcel allocations (for a prespecified number of items per parcel, parcels per factor, and items per factor). This parcel-allocation variability in $\hat{F}_p$ arises because alternate $\mathbf{A}$ matrices (i.e., alternate item-to-parcel-allocations) within a single sample in Equation 5 will result in alternate $\mathbf{\Lambda}_p$ and $\mathbf{\Psi}_p$ across allocations, which will then result in alternate contributions of $\mathbf{C}_{uu_i}$, $\mathbf{C}_{cu_i}$, and $\mathbf{C}_{uc_i}$ to $\hat{F}_p$ across allocations. (This dependency of $\hat{F}_p$ on the particular allocation chosen means that, across allocations within sample, one could draw meaningfully different conclusions about a parcel-solution's model fit [e.g., good fit vs. poor fit], as well as meaningfully different conclusions about point estimates and inferences for structural and measurement parameters—even despite unidimensional, normal items and no model error [Sterba & MacCallum, 2010].)

Putting together the points made by Bandalos (2002) and Sterba and MacCallum (2010) regarding Mechanisms 2 and 3, there should be a relationship between the magnitude of the *typical* within-sample $(\hat{F}_i - \hat{F}_p)$ difference and the magnitude of within-sample $\hat{F}_p$ variability. Consider the fact that, within a given sample, there is only one item-solution, but there are as many parcel-solutions as there are alternative item-to-parcel allocations made—say, for instance, 100. Hence, within a given sample, we can obtain one $\hat{F}_i$ and 100 difference scores of $(\hat{F}_i - \hat{F}_p)$,[6] as well as the across-allocation variance of $\hat{F}_p$. We designate a "typical" item–parcel difference in the minimized discrepancy function value as the average of the 100 difference scores across allocations within sample, or $E_a(\hat{F}_i - \hat{F}_p)$. We denote the variance of $\hat{F}_p$ across allocations within sample as $VAR_a(\hat{F}_p)$. Because there is only one item-solution per sample, $VAR_a(\hat{F}_p) = VAR_a(\hat{F}_i - \hat{F}_p)$. $E_a(\hat{F}_i - \hat{F}_p)$ will be independent of $VAR_a(\hat{F}_i - \hat{F}_p)$, and thus independent from $VAR_a(\hat{F}_p)$, when $\hat{F}_i - \hat{F}_p$ is normally distributed. Discrepancy functions are right-skewed, however, and thus we could expect that almost always the $E_a(\hat{F}_i - \hat{F}_p)$ for a given sample will be positively and monotonically related to $VAR_a(\hat{F}_p)$ for that sample. But because an exact analytic expression for the relationship between $E_a(\hat{F}_i - \hat{F}_p)$ and $VAR_a(\hat{F}_p)$ is not feasible, a simulation demonstration is necessary.

Hypothesis

The previous section leads to the hypothesis that within-sample average $(\hat{F}_i - \hat{F}_p)$ differences should be positively and monotonically related to within-sample parcel-allocation variability in $\hat{F}_p$. Hence, we should be able to predict the amount of $(\hat{F}_i - \hat{F}_p)$ on average, with the amount of parcel allocation variability observed in $\hat{F}_p$ (again, for a fixed number of items per factor, items per parcel, and parcels per factor). The strength of this relationship could vary from sample

[6]Bandalos (2002) and Sass and Smith (2006) also allowed that there would be a different $(\hat{F}_i - \hat{F}_p)$ for different allocations, but did not state the implications of this.

to sample based on sample characteristics, and so should be most pronounced when averaging across samples. If this hypothesis is supported, one implication is that improvements in fit from the item- to parcel-solutions (that many researchers might view as a benefit) should on average be associated with greater parcel-allocation variability in the parcel-solution (a cost)—to a degree dependent on what factors the fit index incorporates other than the discrepancy function value.

There is an important precondition for Mechanisms 2 and 3 from the previous section to create item–parcel fit differences, or parcel-allocation variability in fit; this precondition must be met to test our hypothesis. This precondition is the existence of some amount of error, which can then be reduced or repackaged via parceling, to an allocation-specific extent. This error can take the form of sampling error or measurement model error, or both. Accordingly, parcel-allocation variability in fit has recently been found to be higher when there is more error of either kind in the measurement model (Sterba & MacCallum, 2010). In a separate literature, item–parcel differences in fit have generally been found to be higher when there is more error of either kind in the measurement model (e.g., Bandalos, 2002; Hau & Marsh, 2004; Nasser & Wisenbaker, 2003). Because parceling in general, and certainly the use of parceling in hopes of obtaining improved fit in particular, have been strongly discouraged in the context of measurement model misspecifications, this hypothesis is tested in a sterile context that would not violate this recommendation: where sampling error is present, but model error is not, and furthermore, items are unidimensional and normal in the population. Sampling error is amplified most directly by lowering sample size, but also by having fewer items per fewer parcels, and lower communalities. This sterile context is not a necessity for testing our hypothesis, but it serves to illustrate our points in the simplest manner.

Study 1 tests the hypothesis about the relationship between typical $(\hat{F}_i - \hat{F}_p)$ and parcel-allocation variability in $\hat{F}_p$ using a simulation. Later, Study 2 shows that the relationship between average item–parcel differences in fit indexes and parcel-allocation variability in fit indexes is more complicated than when working directly with the discrepancy function value, as in Study 1.

Methods

To test the hypothesis, a simulation was conducted. This simulation design was used to study parcel-solutions only—not item-solutions, nor their relation—in Sterba and MacCallum (2010). Normally distributed item-level data were generated from a correlated two-factor confirmatory factor population model with either $h = 9$ items per factor or $h = 15$ items per factor.[7] Within each factor items were unidimensional in the population. When generating item-level data, several conditions were varied to elicit a range of high to low sampling error—and consequently high to low parcel-allocation variability in $\hat{F}_p$ and high to low magnitude of $(\hat{F}_i - \hat{F}_p)$—across which to test the hypothesis.

Sample sizes were $N = 75$, 100, 125, 150, 200, or 250; these Ns are commonly employed in SEM research (Baumgartner & Homburg, 1996; Hulland, Chow, & Lam, 1996; MacCallum

[7]Others have considered the generating model to be at the parcel-level, and conceive of it as being mis- or properly specified depending on the sample allocation chosen (e.g., Kim & Hagtvet, 2003), but I believe most researchers conceptualize their generating model at the item-level.

& Austin, 2000). Additionally, each factor had either $j = 3$ parcels with 3 items per parcel (using the $h = 9$ items per factor generated data), or $j = 5$ parcels with 3 items per parcel (using the $h = 15$ items per factor generated data), or $j = 3$ parcels with 5 items per parcel (using the $h = 15$ items per factor generated data). These different combinations of number of items and number of parcels invoke Mechanism 1 from an earlier section (i.e., change in *df* and matrix dimensionality from item- to parcel-solution) to different extents.

Factor variances were both 1, the factor correlation was .25, and item loadings were: high (all .7), medium-high (all .6), low (all .4), or medium-mixed (.4, .5, and .6). These item loadings were chosen to imply a range of scale reliabilities for the factor that are seen in practice (Nunnally & Bernstein, 1994)—from excellent (.90–94, depending on number of items per factor), for high loadings, to good (.84–.89), for medium loadings, to fair (.63–.74), for low loadings. Error variances were chosen to make item variances equal 1.

This design resulted in $6 \times 3 \times 4 = 72$ cells. In each design cell, 100 samples of items were generated. In each sample, 100 random item-to-parcel allocations were performed by randomly assigning the h items per factor to j parcels per factor, where h, j, and numbers of items per parcel were fixed across allocations within a design cell. Then, parcel indicator scores were computed by averaging these randomly assigned items.

For each sample in a cell, an item-solution was obtained by fitting the population-generating two-factor CFA model to that sample's item-level dataset. For each sample in a cell, 100 parcel-solutions were obtained by fitting a two-factor CFA to the 100 differently allocated parcel datasets. M*plus* software (L. K. Muthén & Muthén, 1998–2008) with maximum likelihood estimation was used for model fitting. The discrepancy function value at the final parameter estimates (for Study 1) along with a variety of fit indexes (for Study 2) were recorded for each item-solution and each parcel-solution within a given sample. Improper solutions were removed prior to data analysis; their removal versus inclusion did not change the pattern of findings.

Results

Our measure of the magnitude of the difference between $\hat{F}_i$ and $\hat{F}_p$ is the sample average of the 100 raw $(\hat{F}_i - \hat{F}_p)$ difference scores per sample. Our measure of the amount of parcel-allocation variability in $\hat{F}_p$ is the standard deviation of the across-allocation distribution of $\hat{F}_p$ within a sample. In practice, if a researcher were relying on a single parcel allocation in his or her given sample, the range of the parcel-allocation distribution would also be relevant. Therefore, for perspective, note that for a cell with a medium-sized difference in $\hat{F}_i - \hat{F}_p$ ($N = 150$, 3 items per parcel, 5 parcels per factor, medium loadings), the range of a typical sample's $\hat{F}_p$ allocation distribution is about five times its reported standard deviation.

We begin by investigating the association between parcel-allocation variability in $\hat{F}_p$, and the raw $(\hat{F}_i - \hat{F}_p)$ difference, on average—that is, averaging both quantities across all samples within each cell, in Figure 1. Subsequently we investigate this association at the sample level. In Figure 1, each dot is a cell average; the color of the dots corresponds to a particular number of items and number of parcels (black = 3 items per each of 5 parcels per factor; gray = 5 items per each of 3 parcels per factor; white = 3 items per each of 3 parcels per factor); and the size of the dot corresponds to sample size (largest dot: $N = 250$; smallest

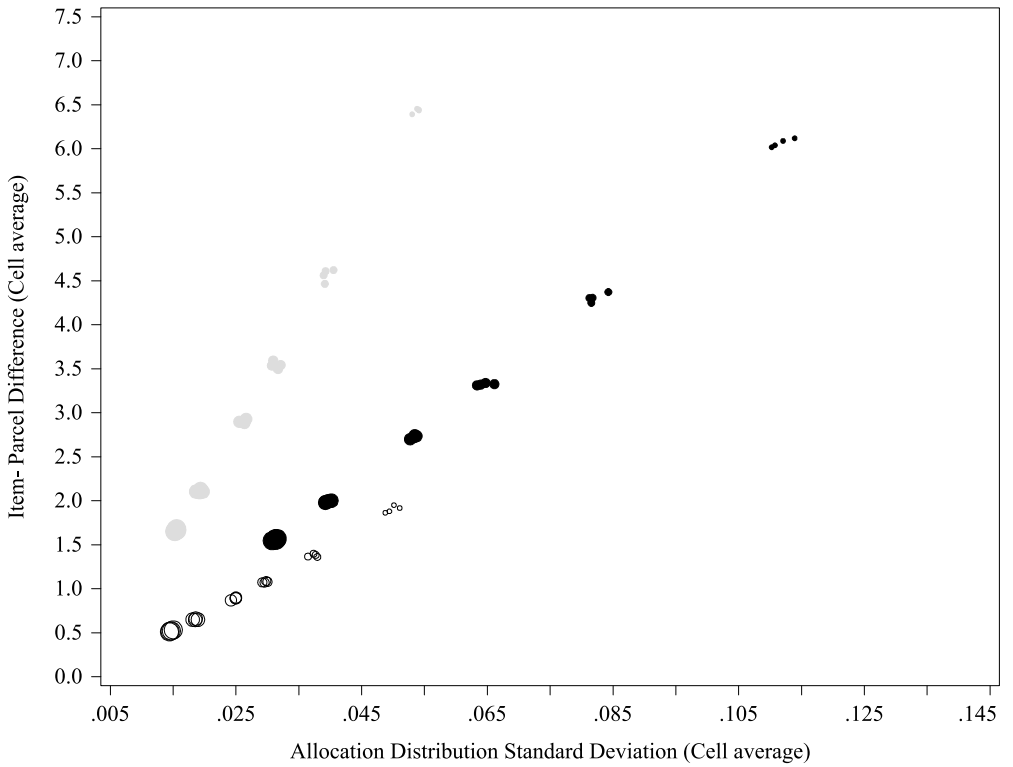

FIGURE 1 Relationship between the cell-average item- versus parcel-solution difference in the discrepancy function value and the cell-average magnitude of parcel-allocation variability in the discrepancy function value. *Note.* Each dot corresponds with a cell average. The color of the dot represents the number of items per factor, parcels per factor, and items per parcel. Specifically, black = 15 items/factor, 5 parcels/factor, and 3 items/parcel. Gray = 15 items/factor, 3 parcels/factor, and 5 items/parcel. White = 9 items/factor, 3 parcels/factor, and 3 items/parcel. The *size* of the dot is proportional to sample size. The smallest dot is $N = 75$; one size larger dot is $N = 100$; one size larger dot is $N = 125$; one size larger dot is $N = 150$; one size larger dot is $N = 200$, and the largest dot size is $N = 250$. Each cluster of four same-size, same-color dots corresponds with the four different loading conditions, but these are not labeled in the plot (see text for explanation).

dot: $N = 75$). Within each cluster of same-size, same-color dots, the four dots correspond to different loading sizes. Because they performed equivalently here, these loading sizes are not labeled to reduce clutter. Figure 1 tells us that, for a given number of items per number of parcels (i.e., a given color), across levels of N there is an extremely strong positive relationship ($r > .99$) between the average amount of allocation-to-allocation variability in the minimized parcel-solution discrepancy function and the average item–parcel difference in the minimized discrepancy function, as hypothesized. More error in the measurement model (here sampling error, which is mainly being driven by N) corresponds on average with both larger item–parcel differences and larger parcel-allocation variability.

The next thing to notice about Figure 1 is that the slope of the relationship between item–parcel differences in the discrepancy function and allocation-variability in the discrepancy

function depends on the number of items per number of parcels. For instance, the gray dots have the steepest slope largely because they show the biggest drop in number of indicators from the item-solution (15 per factor) to the parcel-solution (3 per factor). The black dots have the next steepest slope and the next biggest drop in number of indicators from the item-solution (15 per factor) to the parcel-solution (5 per factor). The white dots show the shallowest slope and also the smallest drop in indicators: 9 per factor in the item-solution to 3 per factor in the parcel-solution.

Also notice that for each combination of number of items per number of parcels (i.e., each color) in Figure 1, the implied intercept would occur where *both* average parcel-allocation variability and average item–parcel differences are nil. In this special case of no model error, this implied intercept would occur when sampling error is nil (i.e., at arbitrarily large N) such that both the item- and parcel-solutions fit perfectly (consistent with Sterba & MacCallum, 2010). If model error was present, arbitrarily large N could not reduce average parcel-allocation variability or average item–parcel differences to nil.

Of course, Figure 1 describes the association between item–parcel differences and parcel-allocation variability in the minimized discrepancy function value, collapsing across all samples. For a given sample, the relationship between a typical item–parcel difference and the magnitude of allocation variability might deviate from its cell average. To visually depict how this association varies from sample to sample, Figure 2 plots 90% confidence ellipses for the cells with 3 items per each of 5 parcels per factor (black ellipses); the cells with 5 items per each of 3 parcels per factor (gray ellipses); and the cells with 3 items per each of 3 parcels per factor (dashed ellipses). An ellipse contains approximately 90% of samples in its designated cell. The plus signs correspond with the cell means from Figure 1. Hierarchical linear regressions, shown in Table 1, quantify the results in Figure 2: The majority (i.e., 88%–95%) of the variance in a given sample's average item–parcel difference in the minimized discrepancy function can be explained by the standard deviation of that sample's parcel-allocation distribution. Figure 2 suggests that the predictive ability of the parcel-allocation standard deviation is somewhat weaker at small sample sizes (i.e., wider ellipses at the far right of the plots). This is conveyed in Table 1 by the 1% to 3% of additional variance explained by N (most for the condition with fewest items per fewest parcels, because it induces most sampling error), and the 2% to 4% of additional variance explained by the interaction of N and the standard deviation of the parcel-allocation distribution. Additional predictors (e.g., factor loadings) explain trivial proportions of variance.[8] Total R^2s ranged from .92 to .98.

Summary

In sum, the findings from Study 1 support the hypothesis that was derived from the theoretical framework presented earlier. That is, the larger the average gap between the item-solution's and parcel-solution's discrepancy function value, the more the parcel-solution's discrepancy function value tends to vary from allocation to allocation—and thus the more it depends on the particular allocation chosen. This association was less strong at low N. Of course, researchers

[8] Additional higher order interaction and power terms involving the predictors were tried in the Table 1 regression model as well as Study 2 regression models, but because they did not explain sizable amounts of variance they were not included for parsimony.

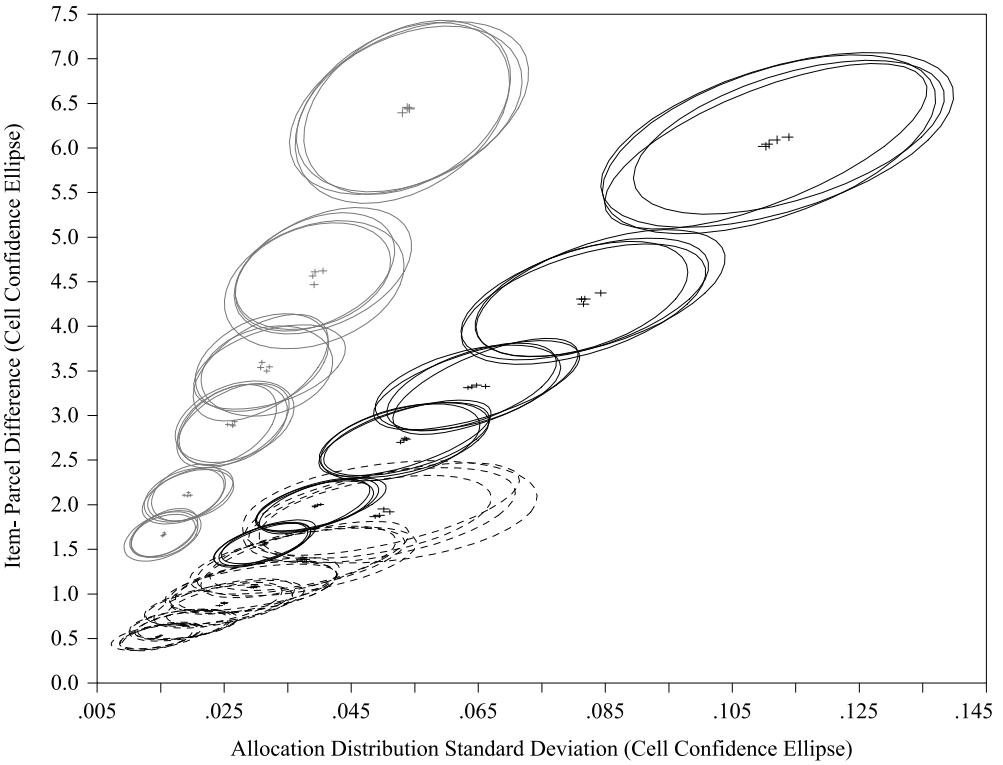

FIGURE 2 Relationship between the sample-average item- versus parcel-solution difference in the discrepancy function value and the sample magnitude of parcel-allocation variability in the discrepancy function value. *Note*. Cell averages from Figure 1 are now denoted by plus signs, and the ellipse around the cell average contains 90% of the sample estimates for that cell. Black ellipses = 15 items/factor, 5 parcels/factor, and 3 items/parcel. Gray ellipses = 15 items/factor, 3 parcels/factor, and 5 items/parcel. Dashed ellipses = 9 items/factor, 3 parcels/factor, and 3 items/parcel.

rarely examine the discrepancy function value itself; instead, they examine model fit indexes that are usually based on the discrepancy function value—the topic of Study 2.

STUDY 2: RELATIONSHIP OF AVERAGE ITEM-PARCEL FIT DIFFERENCES TO PARCEL-ALLOCATION VARIABILITY IN FIT

At a general level, prior literature has typically found improvements in fit from the item-solution to the parcel-solution (e.g., Bandalos, 2002; Hau & Marsh, 2004; Nasser & Wisenbaker, 2003). However, at a more detailed level, the magnitude of such improvements has been found to differ according to the particular model fit index used, with some model fit indexes seemingly more sensitive than others to differences between the item- and parcel-solution. Additionally, the parcel-solution has not always been found to be better fitting than the item-solution for certain fit indexes, particularly in the case of no model error and larger samples (e.g., for root mean

TABLE 1
Dependent Variable = Sample Average Item–Parcel Difference in the Discrepancy Function Value

	3 Items Per 5 Parcels/Factor (Solid Black in Figure 2)					*5 Items Per 3 Parcels/Factor (Solid Gray in Figure 2)*					*3 Items Per 3 Parcels/Factor (Dashed in Figure 2)*				
	Est.	*SE*	*t Value*	ΔR^2	*Cum.* R^2	*Est.*	*SE*	*t Value*	ΔR^2	*Cum.* R^2	*Est.*	*SE*	*t Value*	ΔR^2	*Cum.* R^2
Intercept	2.020	.061	33.120		.000	3.079	.068	45.360		.000	.856	.026	32.800		.000
Parcel-allocation *SD*	58.382	.592	98.580	**.950**	.950	120.355	1.624	74.100	**.880**	.880	34.241	.701	48.850	**.834**	.834
N	<.001	<.001	1.460	.006	.956	−.001	<.001	−5.160	.038	.918	−.001	1	−6.030	.051	.885
Low loadings	.015	.035	.430	<.001	.956	.017	.046	.360	<.001	.918	−.032	.020	−1.620	<.001	.885
Mixed loadings	−.003	.035	−.100	<.001	.956	.062	.046	1.340	<.001	.918	−.054	.019	−2.790	<.001	.885
Medium loadings	.029	.035	.830	<.001	.956	.040	.046	.880	<.001	.918	−.020	.019	−1.010	<.001	.885
Parcel-allocation *SD* * *N*	−.305	.007	−42.670	.019	.975	−.773	.016	−49.850	.042	.960	−.182	.006	−29.750	.031	.916
Parcel-allocation *SD* * Low load	−.230	.502	−.460	<.001	.975	−.340	1.379	−.250	<.001	.960	1.183	.613	1.930	<.001	.916
Parcel-allocation *SD* * Mixed load	.134	.501	.270	<.001	.975	−2.516	1.367	−1.840	<.001	.960	2.230	.594	3.750	<.001	.917
Parcel-allocation *SD* * Medium load	−.393	.494	−.800	<.001	.975	−1.724	1.359	−1.270	<.001	.960	.934	.601	1.550	<.001	.917

Note. Cum. = Cumulative.

square error of approximation [RMSEA]; Bandalos, 2002; Nasser & Wisenbaker, 2003). There has been no integrated explanation of these complexities; rather, these findings have been previously tied together under the expectation that fit will generally improve with parceling.

A relevant point to recall here is that, for a given number of items per number of parcels, the second and third mechanisms identified by the theoretical framework induce item–parcel differences in the discrepancy function values ($\hat{F}_i$ vs. $\hat{F}_p$)—not item–parcel differences in the values of a fit index directly. Likewise, those same mechanisms induce parcel-allocation variability in $\hat{F}_p$—not parcel-allocation variability in the values of a fit index directly. Therefore, any factors that enter into the computation of a model fit index above and beyond the discrepancy function have the potential to modulate (a) the extent to which the item-solution and parcel-solution fit differs, as well as (b) the amount of parcel-allocation variability in that fit index. Study 2 aims to provide some additional insight into the behavior of item vs. parcel fit differences by showing to what extent they can be predicted by allocation-to-allocation variability in that fit index, and to what extent such prediction is affected by index-specific factors.

First consider Table 2, which juxtaposes typical item–parcel fit differences against parcel-allocation distribution ranges for several commonly used fit indexes: RMSEA, standardized root mean squared residual (SRMR), Tucker–Lewis Index (TLI), comparative fit index (CFI), and the chi-square statistic (χ^2). For SRMR, RMSEA, TLI, and CFI, the range of parcel-solution fit from one allocation to another within a sample tends to be *as large as or larger* than the average item–parcel difference in fit for that sample; for χ^2, the same applies to its p value rather than the statistic itself. Consequently, when the difference between item-solution fit and parcel-solution fit is large enough for conclusions about model fit to flip from poor (e.g., TLI $<$ .95, RMSEA $>$ .06; SRMR $>$ .08; χ^2 p value $<$ .05; CFI $<$ .95) to good (e.g., TLI $\geq$ .95, RMSEA $\leq$ .06; SRMR $\leq$ .08; χ^2 p value $\geq$ .05; CFI $\geq$ .95) between the item- and parcel-solutions, parcel-allocation variability tends to be large enough to cause model fit to flip from poor to good across alternative allocations for a parcel-solution.[9]

To further illustrate the relationship between item–parcel fit differences and parcel-allocation variability in fit, Table 3 presents the correlation between cell average item–parcel fit differences and cell average parcel-allocation standard deviation for a given number of items and number of parcels; it is often, but not always $>$ .90. The relationship in a given sample, however, can of course depart from its cell average. Hence, also consider Tables 4 through 7, where for some fit indexes (CFI, TLI, RMSEA) most of the variability in sample-average item–parcel fit differences is predicted by the standard deviation of the corresponding sample's parcel-allocation distribution, but not for other fit indexes (SRMR, χ^2).

Table 3 results differ from Figure 1, and Table 4 through 7 results differ from Table 1 because these fit indexes now involve the discrepancy function value in complex, sometimes nonlinear ways, as well as involve additional quantities. For instance, some of these fit indexes (the incremental fit indexes CFI and TLI) are nonlinear functions of not only the discrepancy function for the model of interest, its *df*, and N, but also of the discrepancy function and *df* for a baseline model (that usually just involves estimating measured variable variances). Other of these fit indexes have censored distributions (e.g., RMSEA and CFI). Still other fit indexes (SRMR) are solely a function of residuals, which constitute only part of the discrepancy function.

[9]Cutoff values are often used to distinguish good fit from poor fit in SEM, but the particular choice of cutoff value for a given index is ultimately arbitrary. Other choices for cutoff values could instead be used to make the same point (e.g., RMSEA $<$ .05 and SRMR $<$ .06 as designating good fit).

TABLE 2
Sample Average Item- vs. Parcel-Solution Fit Differences and Parcel-Allocation Fit Distribution Ranges

		TLI		*CFI*		χ^2 *(p Value)*		*SRMR*		*RMSEA*	
i/p, p/f	*N*	*Item–Parcel Difference*	*Parcel-Allocation Range*	*Item–Parcel Difference*	*Parcel-Allocation Range*	*Item–Parcel Difference*	*Parcel-Allocation Range*	*Item–Parcel Difference*	*Parcel-Allocation Range*	*Item–Parcel Difference*[a]	*Parcel-Allocation Range*
3, 5	75	.169	.212	.149	.093	455.06 (.39)	42.50 (.97)	.028	.042	.032	.103
	100	.099	.155	.086	.069	430.89 (.37)	41.26 (.96)	.024	.037	.025	.087
	125	.068	.121	.057	.053	415.63 (.36)	40.75 (.97)	.022	.032	.021	.076
	150	.050	.099	.041	.044	409.00 (.34)	39.97 (.96)	.020	.030	.018	.068
	200	.032	.075	.025	.033	398.09 (.31)	39.63 (.97)	.017	.026	.015	.059
	250	.024	.059	.017	.026	390.20 (.30)	38.89 (.96)	.015	.023	.013	.051
5, 3	75	.175	.204	.156	.070	482.47 (.44)	20.10 (.97)	.045	.059	.042	.150
	100	.106	.152	.092	.052	456.57 (.42)	19.80 (.97)	.039	.050	.034	.128
	125	.070	.119	.059	.042	442.94 (.39)	19.60 (.97)	.035	.045	.029	.114
	150	.054	.100	.044	.035	435.40 (.37)	19.57 (.97)	.032	.041	.026	.104
	200	.033	.074	.024	.026	422.45 (.34)	19.03 (.97)	.028	.035	.021	.088
	250	.024	.059	.016	.021	416.08 (.33)	19.24 (.97)	.025	.032	.018	.079
3, 3	75	.111	.362	.071	.102	142.82 (.33)	18.32 (.96)	.035	.058	.036	.140
	100	.075	.238	.045	.077	137.90 (.31)	18.36 (.96)	.030	.050	.029	.121
	125	.057	.185	.033	.060	134.88 (.30)	18.07 (.96)	.027	.045	.026	.107
	150	.048	.158	.026	.052	133.47 (.30)	18.21 (.96)	.025	.041	.023	.098
	200	.033	.118	.016	.040	130.58 (.29)	18.10 (.96)	.021	.036	.020	.085
	250	.028	.094	.014	.031	130.54 (.30)	18.08 (.96)	.019	.032	.018	.075

Note. This table contains within-sample quantities that have been marginalized across samples/cell and over cells with different loading sizes. i/p = #items/parcel; p/f = #parcels/factor; TLI = Tucker–Lewis Index; CFI = Comparative Fit Index; SRMR = standardized root mean squared residual; RMSEA = root mean squared error of approximation.

[a]For RMSEA, this is an absolute difference for reasons discussed subsequently in the article.

TABLE 3
Correlation Between the (Cell Average) Difference in Fit Between Item- and Parcel-Solutions and the (Cell Average) Standard Deviation of That Fit Index's Parcel-Allocation Distribution

Fit Index	*Items/Parcel, Parcels/Factor*	*Correlation*
RMSEA	3, 5	.996
	5, 3	.997
	3, 3	.996
CFI	3, 5	.971
	5, 3	.970
	3, 3	.980
SRMR	3, 5	.874
	5, 3	.870
	3, 3	.948
TLI	3, 5	.971
	5, 3	.976
	3, 3	.993
χ^2	3, 5	.921
	5, 3	.723
	3, 3	.491

Note. For RMSEA only, differences in fit are absolute not raw, for reasons discussed subsequently in the text. TLI = Tucker–Lewis Index; CFI = Comparative Fit Index; SRMR = standardized root mean squared residual; RMSEA = root mean square error of approximation.

Such features of fit indexes affect the relationship between item–parcel difference and parcel-allocation variability, largely making it less strong than was the case for the discrepancy function values in Study 1 (i.e., Figure 1 and Table 1). Space does not permit a detailed discussion of how each fit index's particular formulation and its additional terms besides the discrepancy function value (e.g., *df*, *N*) affect this relationship. Instead, we provide such a discussion for two example fit indexes—one with the strongest observed relationship at the sample level between item–parcel fit differences and the parcel-allocation distribution standard deviation (RMSEA), and one with the weakest such relationship (χ^2).

RMSEA

The RMSEA (Steiger & Lind, 1980) measures misfit per *df* in the population. Its sample estimate incorporates several features beyond the discrepancy function value: *N*, *df*, and a max operator:[10]

$$RMSEA = \sqrt{Max\left(\frac{\frac{\hat{F}(N-1)}{df} - 1}{N-1}, 0\right)} \tag{7}$$

[10]M*plus* software employed in this simulation uses N rather than $N - 1$.

TABLE 4
Dependent Variable = Sample Average Absolute Item–Parcel Difference in Root Mean Square Error of Approximation

	3 Items Per 5 Parcels/Factor					*5 Items Per 3 Parcels/Factor*					*3 Items Per 3 Parcels/Factor*				
	Est.	*SE*	*t Value*	ΔR^2	*Cum.* R^2	*Est.*	*SE*	*t Value*	ΔR^2	*Cum.* R^2	*Est.*	*SE*	*t Value*	ΔR^2	*Cum.* R^2
Intercept	.001	.001	1.140		.000	.008	.001	7.050		.000	−.008	.001	−7.250		.000
Parcel-allocation *SD*	1.046	.037	28.440	**.820**	.820	.852	.032	26.960	**.816**	.816	1.054	.034	31.320	**.811**	.811
N	<.001	<.001	−1.850	.006	.826	<.001	<.001	−6.630	.019	.835	<.001	<.001	−.380	.006	.817
Low loadings	−.001	.001	−1.440	<.001	.826	.001	.001	1.010	<.001	.835	−.001	.001	−1.480	<.001	.817
Mixed loadings	−.001	.001	−1.310	<.001	.827	.001	.001	.660	<.001	.835	<.001	.001	.080	<.001	.817
Medium loadings	<.001	.001	−.140	<.001	.827	<.001	.001	−.100	<.001	.835	−.001	.001	−.980	<.001	.817
Parcel-allocation *SD* * *N*	−.001	<.001	−2.590	<.001	.827	<.001	<.001	.160	<.001	.835	.001	<.001	3.830	.001	.818
Parcel-allocation *SD* * Low load	.058	.031	1.860	<.001	.827	−.031	.027	−1.130	<.001	.835	.021	.030	.720	<.001	.818
Parcel-allocation *SD* * Mixed load	.056	.031	1.800	<.001	.828	−.030	.027	−1.120	<.001	.835	−.016	.029	−.540	<.001	.819
Parcel-allocation *SD* * Medium load	.014	.031	.440	<.001	.828	−.008	.027	−.320	<.001	.835	.013	.029	.440	<.001	.819

Note. Cum. = Cumulative.

TABLE 5
Dependent Variable = Sample Average Item–Parcel Difference in Standardized Root Mean Squared Residual

	3 Items Per 5 Parcels/Factor					*5 Items Per 3 Parcels/Factor*					*3 Items Per 3 Parcels/Factor*				
	Est.	*SE*	*t Value*	ΔR^2	*Cum.* R^2	*Est.*	*SE*	*t Value*	ΔR^2	*Cum.* R^2	*Est.*	*SE*	*t Value*	ΔR^2	*Cum.* R^2
Intercept	.025	.000	50.240		.000	.032	.001	46.320		.000	.026	.001	35.050		.000
Parcel-allocation *SD*	1.483	.072	20.590	**.282**	.282	2.164	.077	28.060	**.397**	.397	1.391	.077	18.020	**.352**	.352
N	<.001	<.001	−6.430	.477	.759	<.001	<.001	−5.660	.439	.836	<.001	<.001	−4.350	.380	.732
Low loadings	.004	<.001	10.190	.045	.804	.013	.001	22.700	<.001	.836	.007	.001	10.720	.001	.733
Mixed loadings	.005	<.001	11.550	.012	.816	.011	.001	20.130	.001	.836	.006	.001	11.040	.003	.736
Medium loadings	.003	<.001	7.450	<.001	.816	.006	.001	11.940	.006	.842	.004	.001	6.050	.006	.741
Parcel-allocation *SD* * *N*	−.011	<.001	−25.890	.031	.847	−.012	<.001	−30.850	.029	.871	−.009	<.001	−20.210	.031	.773
Parcel-allocation *SD* * Low load	−1.023	.060	−17.050	.008	.855	−1.282	.064	−20.130	.008	.879	−.533	.065	−8.130	.002	.775
Parcel-allocation *SD* * Mixed load	−.871	.059	−14.710	.009	.864	−1.093	.063	−17.410	.010	.888	−.518	.062	−8.330	.005	.780
Parcel-allocation *SD* * Medium load	−.430	.059	−7.260	.003	.867	−.610	.063	−9.660	.004	.892	−.252	.065	−3.890	.001	.781

Note. Cum. = Cumulative.

TABLE 6
Dependent Variable = Sample Average Item–Parcel Difference in Tucker–Lewis Index or Comparative Fit Index

	3 Items Per 5 Parcels/Factor					*5 Items Per 3 Parcels/Factor*					*3 Items Per 3 Parcels/Factor*				
	Est.	*SE*	*t Value*	ΔR^2	*Cum.* R^2	*Est.*	*SE*	*t Value*	ΔR^2	*Cum.* R^2	*Est.*	*SE*	*t Value*	ΔR^2	*Cum.* R^2
Tucker–Lewis Index															
Intercept	−.061	.007	−8.110		.000	−.028	.008	−3.590		.000	.022	.007	3.190		.000
Parcel-allocation *SD*	7.937	.448	17.720	**.771**	.771	6.585	.449	14.670	**.759**	.759	1.466	.291	5.040	**.530**	.530
N	<.001	<.001	7.970	.015	.785	<.001	<.001	5.160	.010	.769	<.001	<.001	−2.430	.028	.557
Low loadings	.045	.008	5.850	.007	.793	.046	.008	6.150	.002	.771	.084	.006	13.730	.061	.619
Mixed loadings	.006	.005	1.120	.003	.796	.015	.006	2.600	.002	.773	.005	.005	.860	.015	.634
Medium loadings	−.002	.005	−.410	<.001	.796	.003	.005	.610	<.001	.773	.002	.005	.370	.002	.635
Parcel-allocation *SD* * *N*	−.019	.002	−11.690	.006	.802	−.026	.002	−15.580	.021	.795	−.007	.001	−7.670	.007	.642
Parcel-allocation *SD* * Low load	−2.835	.392	−7.240	.007	.809	−.806	.389	−2.080	<.001	.795	−.372	.262	−1.420	.013	.655
Parcel-allocation *SD* * Mixed load	−1.592	.372	−4.280	.002	.811	−.541	.375	−1.440	<.001	.795	.499	.254	1.960	.001	.656
Parcel-allocation *SD* * Medium load	−.502	.374	−1.340	<.001	.811	−.111	.383	−.290	<.001	.795	.235	.263	.890	<.001	.656
Comparative Fit Index															
Intercept	−.033	.005	−6.660		.000	.004	.006	.660		.000	−.011	.005	−2.360		.000
Parcel-allocation *SD*	10.267	.547	18.770	**.807**	.807	12.335	.961	12.830	**.704**	.704	3.819	**.524**	7.290	.553	.553
N	<.001	<.001	5.130	.009	.816	.000	.000	.700	.020	.724	<.001	<.001	1.860	.001	.554
Low loadings	−.007	.005	−1.460	.011	.827	.030	.006	5.200	<.001	.724	−.012	.004	−2.790	.010	.564
Mixed loadings	−.007	.004	−1.860	.004	.831	.009	.005	1.940	<.001	.725	−.013	.004	−3.460	.001	.565
Medium loadings	−.005	.004	−1.410	<.001	.832	.003	.005	.720	<.001	.725	−.004	.004	−.950	<.001	.565
Parcel-allocation *SD* * *N*	−.019	.002	−9.400	.004	.836	−.059	.003	−17.340	.036	.761	−.003	.002	−2.070	.001	.566
Parcel-allocation *SD* * Low load	−1.583	.476	−3.320	.002	.838	.938	.839	1.120	<.001	.761	−.063	.469	−.130	.003	.568
Parcel-allocation *SD* * Mixed load	−.850	.471	−1.810	.001	.838	.673	.839	.800	<.001	.761	.754	.474	1.590	.001	.569
Parcel-allocation *SD* * Medium load	.032	.493	.070	<.001	.838	.505	.880	.570	<.001	.761	.264	.512	.520	<.001	.569

Note. Cum. = Cumulative.

TABLE 7
Dependent Variable = Sample Average Item–Parcel Difference in the Chi-Square Statistic

	3 Items Per 5 Parcels/Factor					*5 Items Per 3 Parcels/Factor*					*3 Items Per 3 Parcels/Factor*				
	Est.	*SE*	*t Value*	ΔR^2	*Cum.* R^2	*Est.*	*SE*	*t Value*	ΔR^2	*Cum.* R^2	*Est.*	*SE*	*t Value*	ΔR^2	*Cum.* R^2
Intercept	259.121	14.996	17.280		.000	406.914	13.124	31.010		.000	102.114	4.948	20.640		.000
Parcel-allocation *SD*	24.591	1.821	13.500	**.325**	.325	21.885	3.311	6.610	**.134**	.134	11.331	1.295	8.750	**.257**	.257
N	−.078	.080	−.980	.194	.519	−.286	.066	−4.340	.259	.393	−.039	.026	−1.510	.043	.300
Low loadings	11.312	13.122	.860	<.001	.520	7.033	11.265	.620	<.001	.393	−9.009	4.362	−2.070	<.001	.301
Mixed loadings	19.219	13.301	1.440	<.001	.520	−.397	11.195	−.040	<.001	.393	−8.438	4.206	−2.010	<.001	.301
Medium loadings	12.802	13.417	.950	<.001	.520	−4.788	11.101	−.430	<.001	.393	−3.901	4.243	−.920	<.001	.301
Parcel-allocation *SD* * *N*	−.025	.010	−2.520	.001	.521	−.011	.017	−.640	<.001	.393	−.006	.007	−.900	<.001	.301
Parcel-allocation *SD* * Low load	−1.391	1.612	−.860	<.001	.521	−2.039	2.849	−.720	<.001	.394	2.286	1.150	1.990	<.001	.302
Parcel-allocation *SD* * Mixed load	−2.352	1.639	−1.430	<.001	.521	−.281	2.847	−.100	<.001	.394	2.477	1.106	2.240	.001	.303
Parcel-allocation *SD* * Medium load	−1.465	1.645	−.890	<.001	.521	.826	2.808	.290	<.001	.394	1.139	1.116	1.020	<.001	.303

Note. Cum. = Cumulative.

We next explain how each of these features can affect the prediction of typical item–parcel fit difference ($RMSEA_i - RMSEA_p$) with parcel allocation variability in $RMSEA_p$. For a visual reference, Figure 3 illustrates the RMSEA results from Table 3.

First consider division by *df*. This operation serves to penalize adding estimated parameters unless they meaningfully decrease the discrepancy function value. This penalty for complexity can be loosely thought of as counterbalancing the first mechanism responsible for item–parcel differences in fit—the reduced *df* for the parcel-solution. One observed consequence of the *df* penalty is that the slopes are more similar across each combination of number of items per number of parcels in Figure 3 than in Figure 1. In Figure 3, a .10 difference of ($RMSEA_i - RMSEA_p$) on average is associated with nearly a .10 increase in the $RMSEA_p$'s allocation distribution standard deviation, regardless of the number of items per number of parcels. A second observed consequence of the *df* penalty is to make it likely or plausible for the parcel-solution to possibly fit worse, per *df,* than the item-solution. That is, if a particular amount of misfit in the item-solution remains in the parcel-solution (and was not repackaged into, say, common variance by parceling), it will be weighted more in $RMSEA_p$.

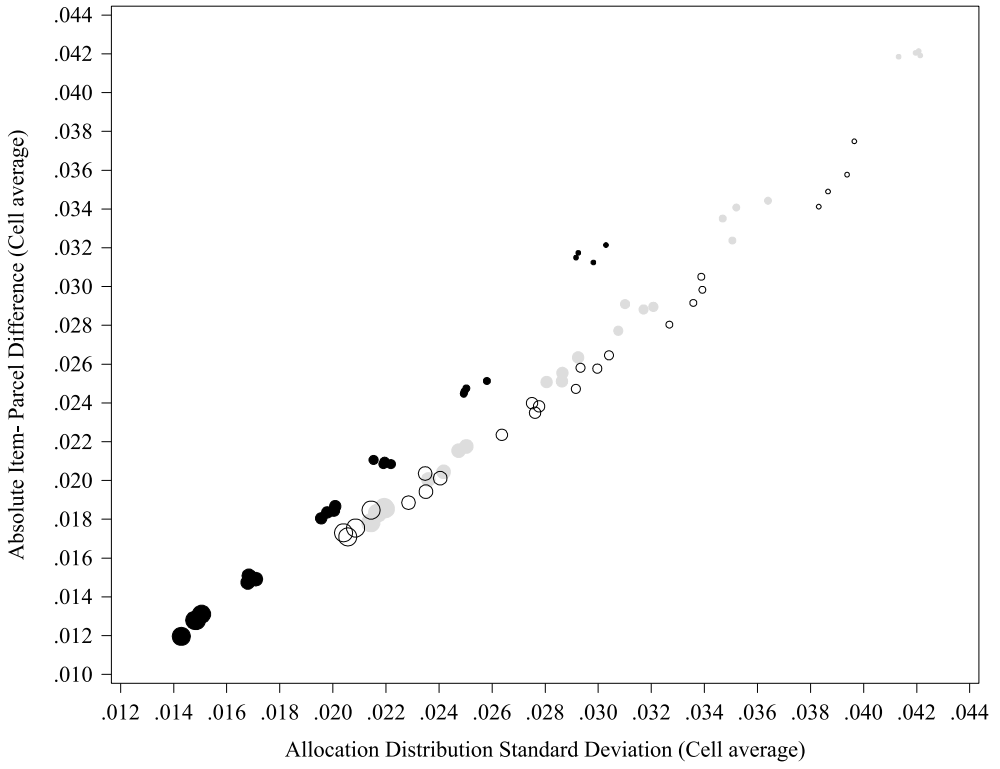

FIGURE 3 Relationship between the cell-average item- versus parcel-solution difference in root mean square error of approximation (RMSEA) and the cell average magnitude of parcel-allocation variability in RMSEA. *Note.* Each dot corresponds with a cell average. The colors and sizes of dots have the same definitions as in Figure 1.

This upweighting of misfit will be due to the parcel-solution's smaller *df*. In contrast to Figure 1, this consequence necessitated consideration of absolute rather than raw item–parcel differences in Figure 3, to capture the net change in $RMSEA_i - RMSEA_p$ and prevent opposite-signed differences from canceling each other out in aggregation. (Raw differences were used for all other indexes.) Approximately one-third of the time $RMSEA_i < RMSEA_p$—although 17% of these instances were inevitable given that $RMSEA_i$ had perfect fit of 0.[11] Likewise, previous studies have occasionally found $RMSEA_i$ to fit better than $RMSEA_p$ (e.g., Bandalos, 2002; Nasser & Wisenbaker, 2003) in the case of no model error.

Next consider N; the manner in which N is included in both the numerator and denominator of the index appears to render the RMSEA no more sensitive to N than the discrepancy function value itself. This can be seen by comparing Figures 1 and 3. Next consider the max operator, which sets negative values of the quantity in large brackets to 0. What constitutes a negative value of the quantity in large brackets itself depends on *df*. This censoring should weaken the association between item–parcel fit difference and parcel-solution allocation variability in particular when either $RMSEA_i$ or $RMSEA_p$ was censored at 0. Whereas the effects of the max operator are not obvious in Figure 3, an exploratory investigation that involved removing versus retaining the max function and radical found that the effects of the left censoring in a given sample served to constrain the partial R^2 for the $RMSEA_p$ allocation standard deviation in Table 3.

Chi-Square

Another commonly used fit index is the chi-square statistic:[12]

$$\chi^2 = \hat{F}(N-1) \tag{8}$$

The direct multiplication of the discrepancy function by $N-1$ served to weaken the prediction of typical item–parcel fit differences (i.e., $\chi_i^2 - \chi_p^2$) with parcel allocation variability in χ_p^2, as seen in Table 3 (also compare Tables 1 and 7). For a visual reference, compare Figure 4, which illustrates the chi-square results from Table 3, to Figure 1, which illustrates the discrepancy function results from Table 3. In Figure 1, the cell means (dots) are inversely related to N, and in Figure 4 the cell means (dots) are relatively insensitive to N. One reason for the differential sensitivity to N between Figures 1 and 4 is as follows. Because $E(\hat{F}) = \frac{df}{N-1}$ in the present context of no model error (Browne & Cudeck, 1993), y-axis values in Figure 1 (i.e., means of differences in discrepancy function values) are inversely related to N. Because $E(\chi^2) = df$ in the present context of no model error (Browne & Cudeck, 1993), y-axis values in Figure 4 (i.e., means of differences in chi-square values) are relatively insensitive to N; rather, they hover around $df_i - df_p$.

[11] For comparison purposes, a subset of cells (24) were rerun with a purposive parceling algorithm—the correlational algorithm from Rogers and Schmitt (2004). This correlational algorithm was designed to outperform random allocations in terms of parcel-solution fit due to its allocation of items that highly correlate into the same parcel (thus reducing the chance of high unaccounted for correlated uniquenesses). Using this algorithm, 23% of item-solutions had better RMSEA fit than parcel-solutions, but 10% of these instances were inevitable due to the item solution having an RMSEA of 0.

[12] M*plus* software employed in this simulation uses N rather than $N-1$.

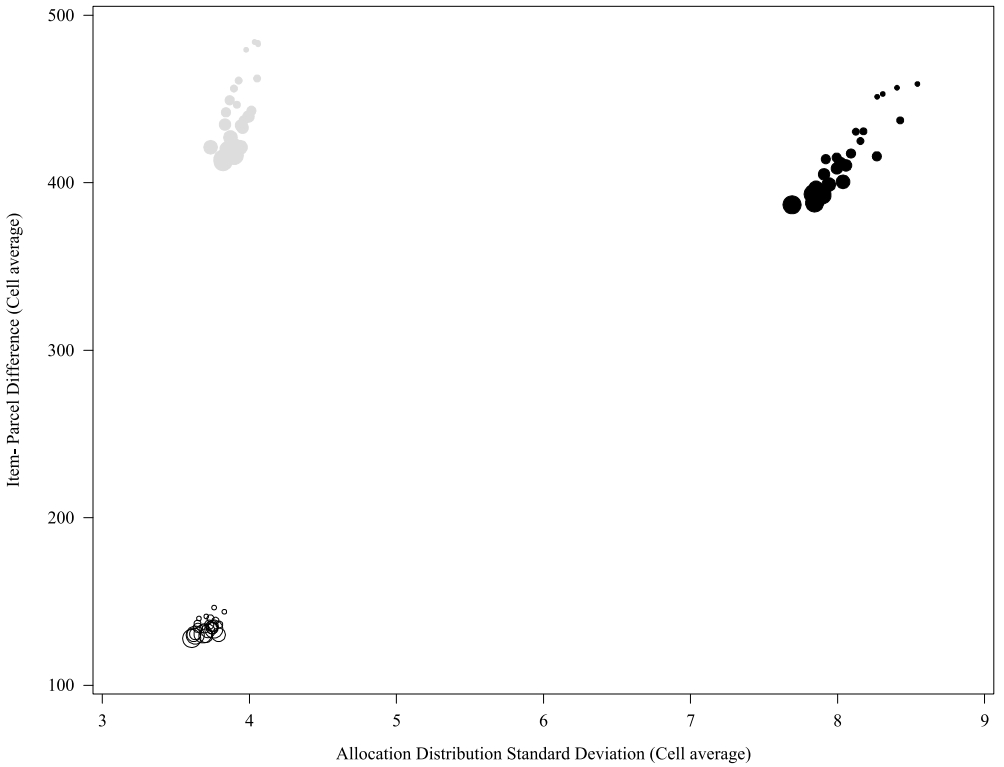

FIGURE 4 Relationship between the cell-average item- versus parcel-solution differences in chi-square and the cell-average magnitude of parcel-allocation variability in chi-square. *Note.* Each dot corresponds with a cell average. The colors and sizes of dots have the same definitions as in Figure 1.

In practice, however, there would be model error, in which case $E(\chi^2) = df + \lambda$, where λ (the noncentrality parameter) is a product of $N-1$ and model error. As such, average $(\chi_i^2 - \chi_p^2)$ would no longer hover around $df_i - df_p$ but would increase with both sample size and model error, which in turn would force the χ_p^2 distribution's standard deviation to do the same. Hence, in the case of model error, we would expect to see stronger positive relationships in Figure 4 for chi-square, more similar to the other fit indexes in Table 3.

Summary

Study 2 extended the results of Study 1 from discrepancy function values to fit indexes. Study 2 illustrated that, for fit indexes, the relationship between average item–parcel differences and parcel-allocation variability was still positive, but varied in strength in ways that could be anticipated based on what elements other than the discrepancy function value (e.g., N or df) entered into the fit index computation, and precisely how they entered the computation. Study 2 illustrated this reasoning for two fit indexes in detail—RMSEA and χ^2. The effect of the additional elements entering the fit index computation is likely one reason why previous

studies have found that some fit indexes appear more sensitive than others to item–parcel fit differences (e.g., Bandalos, 2002; Nasser & Wisenbaker, 2003). The bottom line of Study 2, for an applied researcher, was shown in Table 2: When the typical item–parcel fit difference was large enough to change conclusions about model fit (e.g., from good to poor fit), parcel-allocation variability was on average large enough to change conclusions from good to poor fit across alternate parcel allocations.

GENERAL DISCUSSION

This article provides a crucial link between the many prior studies on item–parcel differences in fit and the single prior study on parcel-allocation variability. It began with a theoretical framework that implied that mechanisms causing item versus parcel differences in discrepancy function values also cause parcel-allocation variability in discrepancy function values. A simulation in Study 1 empirically demonstrated that, indeed, item–parcel differences and parcel-allocation variability in discrepancy function values on average increase nearly in lockstep, for a given number of items per number of parcels. Study 2 showed that the relation between average item–parcel differences in fit and parcel-allocation variability is similarly positive, but varies in nature and magnitude based on how the discrepancy function enters into the index computation, and what other quantities (e.g., N, df) are involved in the index computation. Importantly for applied researchers, when the average item- versus parcel-solution fit difference in a sample was large enough to change substantive conclusions about model adequacy, parcel-allocation variability tended to be large enough for conclusions about model adequacy to depend on the particular allocation chosen (see Table 2). The often-encountered but ill-conceived use of parceling to obtain improved fit should be connected to its associated hidden cost: additional uncertainty interjected into parcel-solution fit indexes in the form of parcel-allocation variability.

Limitations

Several limitations of this study should be considered. First, this study used only random allocations. However, according to the theoretical framework, conclusions should generalize to purposive allocations as well. Second, although the mechanisms causing item–parcel differences and parcel-allocation variability require either the presence of sampling error and/or measurement model error, I included only sampling error. I did so because parceling with the goal of improving fit has already been criticized in the presence of measurement model error due to the potential for obscuring model misspecifications (e.g., Bandalos, 2002; Hall et al., 1999). I wanted to show that problems with the not-infrequent practice of parceling to improve fit were even more far reaching, such that they occur even when best case scenario conditions are upheld. Nonetheless, because some measurement model error is to be expected in the real world, I noted instances in Study 1 and Study 2 where the pattern of the results would be expected to meaningfully change in the presence of model error. Further, I explained what results would be expected if model error were included (e.g., Figure 1 not having nil x-axis and y-axis values at large N; Figure 3 having stronger positive relationships between x-axis and y-axis values). Finally, I used a limited number of random allocations per sample (100).

However, a sensitivity analysis varying the number of allocations in select cells recovered the same overall pattern of findings.

Recommendations

It is suggested that, particularly when item–parcel differences in fit are observed, parcel-allocation variability in model fit should be assessed. Researchers should report whether parcel-allocation variability is large enough to substantively affect conclusions. A utility for assessing parcel-allocation variability is available from http://www.vanderbilt.edu/peabody/sterba (or by contacting the author), and reporting guidelines were provided in Sterba and MacCallum (2010). Although previous studies have condemned using parceling to obtain improved fit based on the potential for parceling to obscure model misspecification (e.g., Bandalos & Finney, 2001; Hall et al., 1999), we showed that—even when there is no measurement model error—item-to-parcel improvements in fit are not the no-strings-attached benefit that Williams and O'Boyle (2008) had in mind. Nontrivial improvements in fit are on average accompanied by the undesirable consequence of nontrivial parcel-allocation variability in fit; hence the latter must be investigated and reported regardless of the existence of model error.

ACKNOWLEDGMENT

I would like to thank Robert C. MacCallum for helpful comments on prior drafts.

REFERENCES

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

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

Bandalos, D. L., & Finney, S. J. (2001). Item parceling issues in structural equation modeling. In G. A. Marcoulides and R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 269–297). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Baumgartner, H., & Homburg, C. (1996). Applications of structural equation modeling in marketing and consumer research: A review. *International Journal of Research in Marketing, 13,* 139–161.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). London, England: Sage.

Hall, R. J., Snell, A. F., & Foust, M. S. (1999). Item parceling strategies in SEM: Investigating the subtle effects of unmodeled secondary constructs. *Organizational Research Methods, 2,* 233–256.

Hau, K.-T., & Marsh, H. W. (2004). The use of item parcels in structural equation modeling: Non-normal data and small sample sizes. *British Journal of Mathematical and Statistical Psychology, 57,* 327–351.

Hulland, J., Chow, Y. H., & Lam, S. (1996). Use of causal models in marketing research: A review. *International Journal of Research in Marketing, 13,* 181–197.

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,* 101–127.

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

Lee, D. (2005). Calls for multiple indices incorporating multiculturalism in content analysis. *The Counseling Psychologist, 33,* 349–357.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Reviews of Psychology, 51,* 201–226.

MacCallum, R. C., & Tucker, L. R (1991). Representing sources of error in the common-factor model: Implications for theory and practice. *Psychological Bulletin, 109,* 502–511.

Martens, M. P. (2005). The use of structural equation modeling in counseling psychology research. *The Counseling Psychologist, 33,* 269–298.

Muthén, L. K., & Muthén, B. O. (1998–2008). *M*plus *user's guide* (5th ed.). Los Angeles, CA: Muthén & Muthén.

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,* 729–757.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

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

Sass, D. A., & Smith, P. L. (2006). The effects of parceling unidimensional scales on structural parameter estimates in structural equation modeling. *Structural Equation Modeling, 13,* 566–586.

Steiger, J. H., & Lind, J. M. (1980, May). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Sterba, S. K., & MacCallum, R. C. (2010). Variability in parameter estimates and model fit across random allocations of items to parcels. *Multivariate Behavioral Research, 45,* 322–358.

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). *Using item "Testlet"/"Parcels" in confirmatory factor analysis: An example using the PPSDQ-78.* (ERIC Document ED 404349)

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