Multivariate Behavioral Research, 45:322–358, 2010
Copyright © Taylor & Francis Group, LLC
ISSN: 0027-3171 print/1532-7906 online
DOI: 10.1080/00273171003680302

Psychology Press
Taylor & Francis Group

Variability in Parameter Estimates and Model Fit Across Repeated Allocations of Items to Parcels

Sonya K. Sterba and Robert C. MacCallum
The University of North Carolina at Chapel Hill

Different random or purposive allocations of items to parcels within a single sample are thought not to alter structural parameter estimates as long as items are unidimensional and congeneric. If, additionally, numbers of items per parcel and parcels per factor are held fixed across allocations, different allocations of items to parcels within a single sample are thought not to meaningfully alter model fit—at least when items are normally distributed. We show analytically that, although these statements hold in the population, they do not necessarily hold in the sample. We show via a simulation that, even under these conservative conditions, the magnitude of within-sample item-to-parcel-allocation variability in structural parameter estimates and model fit can alter substantive conclusions when sampling error is high (e.g., low N, low item communalities, few items per few parcels). We supply a software tool that facilitates reporting and ameliorating the consequences of item-to-parcel-allocation variability. The tool's utility is demonstrated on an empirical example involving the Neuroticism-Extroversion-Openness (NEO) Personality Inventory and the Computer Assisted Panel Study data set.

Parceling—averaging or summing raw items and using the resultant score as a factor indicator—has a long history in psychology (e.g., Cattell, 1956, 1974) and remains quite common. For example, a review of 1998–2001 issues of several

Correspondence concerning this article should be addressed to Sonya K. Sterba, L. L. Thurstone Psychometric Laboratory, UNC-CH CB#3270, Chapel Hill, NC 27599-3270. E-mail: ssterba@email.unc.edu

psychology journals found that 20% of structural equation modeling (SEM) applications used parcel scores as factor indicators (Bandalos & Finney, 2001). Although a variety of methods exist for *allocating* items to parcels (i.e., assigning items to parcels for a prespecified number of parcels/factor and items/parcel), Bandalos and Finney found random or quasi-random[1] parceling methods (45%) to be the most common. In contrast, purposive parceling methods (e.g., allocating items to parcels based on item content or allocating items to parcels based on similar factor loadings in a preliminary exploratory factor analysis) were used in only 15% of studies.[2]

The long-standing use of parceling is motivated by several benefits of parcel-solutions compared with item-solutions. Parceling improves model fit (Bandalos, 2002; Nasser & Wisenbaker, 2003), increases communalities (Bandalos & Finney, 2001; Cattell, 1974), and, for parallel items, increases indicator reliability (Coffman & MacCallum, 2005) compared with item-level solutions. Parceling also ameliorates some effects of coarsely categorized indicators without requiring categorical variable estimation (Bandalos, 2002; West, Finch, & Curran, 1995) and ameliorates some effects of nonnormal continuous indicators (Hau & Marsh, 2004; Nasser & Wisenbaker, 2003; Nasser-Abu & Wisenbaker, 2006) without requiring data transformation, robust normal theory estimation, or distribution-free estimation—some of which require large samples (B. O. Muthén, du Toit, & Spisic, 1997). However, these benefits of parceling are less pronounced under conditions of high item communalities and large sample sizes (e.g., Hau & Marsh, 2004; Marsh, Hau, Balla, & Grayson, 1998). The use of parceling is also subject to several drawbacks (but see counterarguments in Little, Cunningham, Shahar, & Widaman, 2002). Parceling has been shown to potentially obscure unspecified secondary factors or correlated uniquenesses (Hall, Snell, & Foust, 1999, Simulation 2-3) and to potentially obscure measurement variance in multiple group models (Meade & Kroustalis, 2006).

Hence, parceling has been recommended for situations that maximize these benefits and minimize these weaknesses (Little et al., 2002). To maximize benefits, parceling has been particularly recommended for small samples and low item communalities (e.g., Bagozzi & Edwards, 1998; Marsh et al., 1998; Meade & Kroustalis, 2006, p. 372; Nasser-Abu & Wisenbaker, 2006; West et al., 1995). To minimize drawbacks, conservative circumstances for parceling would be (a) a single-group analysis, when (b) there is good a priori knowledge that items to be parceled are, indeed, unidimensional (i.e., no cross loadings, no error covariances, no secondary factors) and congeneric (i.e., each item only loads on one factor) in the population and when (c) the research objective is to study the

[1] Quasi-random allocations could involve parceling adjacent items or even and odd items.

[2] The other 40% of studies did not report the item-to-parcel allocation method used.

structural relationship among latent variables rather than develop scales (e.g., Bandalos & Finney, 2001; Meade & Kroustalis, 2006). This study concerns the use of parceling in this most highly recommended and conservative setting: unidimensional, congeneric items, low sample size, and low item communalities.[3] Moreover, this study uses the most popular method of parceling—random item-to-parcel allocations—to illustrate key points. However, our analytic results hold for purposive item-to-parcel allocations as well (as discussed later). Additionally, this study concerns characteristics of parcel-solutions as compared with each other, not as compared with item-solutions.

Because applied researchers usually make a single (random or purposive) item-to-parcel allocation for a given sample, before using parceling in this most highly recommended setting, one would want to verify the following. One would want assurance that the parcel-solution's structural parameter estimates do not vary a great deal according to the particular item-to-parcel allocation made in that sample, and—at least when the number of parcels/factor and items/parcel is held fixed and items are normally distributed—neither does model fit. We might be assured by prior conclusions that, when items are unidimensional, congeneric, and normally distributed in the population, "it is clear that using item parcels ... will not affect the structural model parameters if the scale is unidimensional" (Sass & Smith, 2006, p. 572). Or, "The use of item parceling had negligible effects on parameter bias and on the standard errors of the estimated factor correlations" (Nasser-Abu & Wisenbaker, 2006, abstract). Or, that "in situations in which a set of items truly measures a single underlying factor, the choice of composite formation strategy should not substantially influence model fit" (Landis, Beal, & Tesluk, 2000, p. 190). Or, that "the choice of item parceling strategy was essentially arbitrary when no secondary influences were present" (Hall et al., 1999, p. 249). These conclusions have been sometimes based on one type of evidence, and sometimes on another.

The first type of evidence contributing to these conclusions involves the *within-allocation sampling distribution* of structural parameter and model fit estimates. An example of a within-allocation sampling distribution, from Nasser and Wisenbaker (2003) and Nasser-Abu and Wisenbaker (2006), is the allocation of 12 items to three 4-item parcels in the same way for 500 samples: Items 1, 2, 11, and 12 to Parcel 1; Items 3, 4, 9, and 10 to Parcel 2; and

[3] Note that throughout this study we assume that the population-generating model of theoretical interest is an item-level model, and parceling is simply used as a tool to facilitate model estimation. This premise lies in contrast to the program of work by Hagtvet, Kim, and colleagues (e.g., Hagtvet & Nasser, 2004; Kim, 2000; Kim & Hagtvet, 2003). They consider there to be one true population-generating *parcel-level* model, which may or may not be misspecified depending on how the researcher allocates items to parcels in his or her sample. We believe our premise is in line with typical motives for parceling in much of psychological research (e.g., Sterba, Egger, & Angold, 2007).

Items 5, 6, 7, and 8 to Parcel 3. Across repeated samples, within this *single,* fixed item-to-parcel allocation of unidimensional/congeneric/normal items, researchers have found little bias in structural parameter estimates (Hau & Marsh, 2004; Marsh et al. 1998; Nasser-Abu & Wisenbaker, 2006), modest variability in model fit (Hau & Marsh, 2004; Marsh et al., 1998; Nasser & Wisenbaker, 2003), and decreasing variability in both with larger N and larger item communalities.

The second type of evidence contributing to these conclusions involves the *within-sample parcel-allocation distribution* of structural parameter and model fit estimates. An example of a within-sample parcel-allocation distribution, from Hall and colleagues (1999, Simulation 1), is the allocation of six items to two three-item parcels in 15 ways for the same sample. Whereas a number of studies have compared alternate within-sample item-to-parcel allocations for simulated multidimensional item sets (Bandalos, 2002; Rogers & Schmitt, 2004), or empirical item sets (where item dimensionality is indeterminate; Kishton & Widaman, 1994; Landis et al., 2000; Sass & Smith, 2006, Study 2), we know of only two studies that have done so for simulated unidimensional, congeneric, normal item sets. Across repeated allocations of unidimensional/congeneric/normal items to parcels within a *single* sample, researchers have found that structural parameter estimates remain unchanged and model fit is nonmeaningfully changed—as long as numbers of parcels/factor and items/parcel are fixed across allocations (Hall and colleagues, 1999, Simulation 1: 15 allocations; Sass & Smith, 2006, Study 1: 2 allocations).

These two sources of evidence are currently used to inform applied researchers regarding the consequences of parceling on structural parameter estimates and model fit for their sample of unidimensional, congeneric items. However, we argue that the first source of evidence is less relevant to the issue at hand and we argue that the second source of evidence has been underinvestigated, possibly leading to unsupported conclusions.

The first source of evidence is practically less relevant to the applied researcher with one sample and many potential ways of allocating items to parcels for that one sample. Such a researcher would likely want to know if Nasser and Wisenbaker (2003) and Nasser-Abu and Wisenbaker (2006) would have obtained similar results for a given sample if they had happened to allocate Items 7, 1, 11, and 3 to Parcel 1; Items 8, 10, 5, and 6 to Parcel 2; and Items 4, 9, 12, and 2 to Parcel 3, or otherwise happened to allocate their 12 items to three 4-item parcels in any of the other $(p!)/(p_1!p_2!p_3!) = 34{,}650$ possible ways.

The second source of evidence is promising. However, because Hall et al. (1999, Study 1) and Sass and Smith (2006, Simulation 1) considered only data conditions with minimal sampling error, the generalizability of Hall et al.'s and Sass and Smith's results to other circumstances is questionable. Both employed large item loadings ($\lambda_i = .65–.85$) and large samples ($N = 500–1{,}000$) and, most important, Hall et al. further minimized sampling error by averaging *across 500 samples* within-allocation before comparing results across their 15 allocations. No studies have investigated the effects of alternative parcel-allocations within a sample of unidimensional/congeneric/normal items when sampling error is nontrivial.

OVERVIEW OF CURRENT STUDY

In this article we first show analytically why, within a single sample, alternate item-to-parcel allocations *can* differentially affect structural parameter estimates—even when (a) items are unidimensional and congeneric and (b) the properly specified item-generating model is fit to parceled data. We also show analytically why, within a single sample, alternate item-to-parcel allocations *can* differentially affect model fit—even if (c) number of items/parcel and parcels/factor are held fixed across allocations and (d) items are normally distributed and normal theory estimation is employed. Our analytic results indicate that the settings showing the most variability in model fit and structural parameter estimates across item-to-parcel allocations should be precisely those settings in which parceling is most recommended—low item communalities and low sample size. That is, within-sample parcel-allocation variability in model fit and structural parameter estimates should depend on sampling error—which is precisely why it remained undetected by Hall et al. (1999, Simulation 1) and Sass and Smith (2006, Study 1), who both minimized sampling error.

Second, we then use a simulation involving random item-to-parcel allocations to provide proof of concept of our analytic results. We also use this simulation to connect the present work on within-sample parcel allocation variability to prior parceling literature on within-allocation sampling variability (e.g., Hau & Marsh, 2004; Marsh et al., 1998; Nasser & Wisenbaker, 2003; Nasser-Abu & Wisenbaker, 2006). Specifically, we compare the amount of within-sample parcel-allocation variability to the amount of within-allocation sampling variability for these outcomes, and show both to be magnified by the same data characteristics.

Third, because our simulation was conducted under the conservative conditions of no model error and unidimensional, congeneric, normal items, we present an empirical illustration of the magnitude of parcel-allocation variability that might be more typically encountered in the real world. The empirical illustration involves the Neuroticism-Extroversion-Openness (NEO) Personality Inventory (Costa & McCrae, 1985) and uses the Computer Assisted Panel Study data set (Latane, 1989). Fourth, we provide an SAS macro that can be used when sampling error is large in order to ameliorate parcel-allocation variability; the macro provides the mean and variability of parameter and model fit estimates across a large distribution of random item-to-parcel allocations.

THEORETICAL FRAMEWORK

Yuan, Bentler, and Kano (1997, Equations 2.1 and 2.2) presented analytic results showing that, when items are unidimensional and congeneric in the population (and other model assumptions are upheld), structural parameters *in the population* will be invariant to the parceling method used (or item-to-parcel allocation made; see also Sass & Smith, 2006, pp. 570–571). Yuan and colleagues' finding is guaranteed to hold only in the population, however, not necessarily in the sample. Because no population model ever holds exactly in the sample (MacCallum, 2003), we present a generalization of Yuan and colleagues' results for the case of population data, and then extend these results to sample data, using the theoretical developments of MacCallum and Tucker (1991) and MacCallum, Widaman, Zhang, and Hong (1999).

Population-Level Data

Suppose that we wish to construct n parcels from a set of m items measuring q factors in the population. Let an i-subscript denote *item level*. Let x_i be a vector of deviation scores on items in the population, of order m, and let E denote the expectation operator, $\mathbf{\Lambda}_i$ denote a $m \times q$ common factor loading matrix, $\mathbf{\Phi}_i$ denote a $q \times q$ covariance matrix of common factors, and $\mathbf{\Psi}_i^2$ denote an $m \times m$ diagonal matrix of unique variances. The population covariance structure of the items can be derived as

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

—under the assumption that the unique factors are uncorrelated with each other and with common factors in the population. In these developments and in the subsequent simulation, we also assume that at the item-level the model fits perfectly in the population, but this assumption could be relaxed (see MacCallum, Widaman, Preacher, & Hong, 2001).

Now let $\mathbf{A}$ be an $m \times n$ selection matrix that allocates items to parcels, *given a prespecified number of parcels/factor and items/parcel*. Let a p-subscript denote *parcel level*. Then $x_p = \mathbf{A}x_i$ is a vector of parcels, of order n. To illustrate, suppose we have $m = 12$ items and $q = 2$ factors, with Items 1–6 loading on the first factor and Items 7–12 loading on the second factor, and suppose we construct two parcels per factor as follows:

$$\underbrace{\begin{bmatrix} x_{p_1} \\ x_{p_2} \\ x_{p_3} \\ x_{p_4} \end{bmatrix}}_{x_p} = \underbrace{\begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_{i_1} \\ x_{i_2} \\ x_{i_3} \\ x_{i_4} \\ x_{i_5} \\ x_{i_6} \\ x_{i_7} \\ x_{i_8} \\ x_{i_9} \\ x_{i_{10}} \\ x_{i_{11}} \\ x_{i_{12}} \end{bmatrix}}_{x_i}$$

$$= \begin{bmatrix} (x_{i_1} + x_{i_2} + x_{i_3})/3 \\ (x_{i_4} + x_{i_5} + x_{i_6})/3 \\ (x_{i_7} + x_{i_8} + x_{i_9})/3 \\ (x_{i_{10}} + x_{i_{11}} + x_{i_{12}})/3 \end{bmatrix}.$$

Here x_{p_1} and x_{p_2} load on the first factor and x_{p_3} and x_{p_4} load on the second factor. Note that it makes no difference here whether the locations of nonzero elements in $\mathbf{A}$ are chosen randomly or purposively. The population covariance matrix of the parcels is then denoted by

$$\boldsymbol{\Sigma}_p = E(x_p x_p'), \tag{2}$$

which can be rewritten as

$$\boldsymbol{\Sigma}_p = E(\mathbf{A} x_i x_i' \mathbf{A}') = \mathbf{A}\boldsymbol{\Sigma}_i \mathbf{A}', \tag{3}$$

which implies that

$$\boldsymbol{\Sigma}_p = \mathbf{A}\boldsymbol{\Lambda}_i \boldsymbol{\Phi}_i \boldsymbol{\Lambda}_i' \mathbf{A}' + \mathbf{A}\boldsymbol{\Psi}_i^2 \mathbf{A}'. \tag{4}$$

Now, rewriting $\boldsymbol{\Lambda}_p = \mathbf{A}\boldsymbol{\Lambda}_i$ and $\boldsymbol{\Psi}_p^2 = \mathbf{A}\boldsymbol{\Psi}_i^2\mathbf{A}'$ we have

$$\boldsymbol{\Sigma}_p = \boldsymbol{\Lambda}_p \boldsymbol{\Phi}_i \boldsymbol{\Lambda}_p' + \boldsymbol{\Psi}_p^2. \tag{5}$$

In other words, the following conditions will hold *in the population* when the factor model fits perfectly at the item level. The factor loading of a parcel (i.e., element of $\boldsymbol{\Lambda}_p$) will equal the average of the factor loadings of its constituent items in an item-level analysis (i.e., constituent elements of $\mathbf{A}\boldsymbol{\Lambda}_i$) when parcels are formed by averaging items.[4] It is important to note that the previous statement refers to comparisons of unstandardized parcel loadings with averaged unstandardized item loadings in the population (not standardized parcel loadings with averaged standardized item loadings). Furthermore, whenever we refer to item-level or parcel-level parameters or estimates in the remainder of these theoretical developments, we are always referring to *unstandardized* parameters or estimates. The unique variance of a parcel (i.e., element of $\mathbf{\Psi}_p^2$) will equal the average of the unique variances of its constituent items in an item-level analysis—divided by the number of items per parcel (i.e., constituent elements of $\mathbf{A}\mathbf{\Psi}_i^2\mathbf{A}'$),[5] when parcels are formed by averaging items.[6] And, the factor covariances (structural parameters) will be the same for a parcel-level or item-level analysis (i.e., $\mathbf{\Phi}_i = \mathbf{\Phi}_p$). Importantly, these results will apply in the population for any $\mathbf{A}$ matrix—that is, for any purposive or random item-to-parcel allocation.

Sample-Level Data

Suppose we draw a sample of observations from the population described earlier and we consider the structure of the item-level sample covariance matrix, $\mathbf{C}_i$. The following developments are based directly on work by MacCallum and Tucker (1991) and MacCallum et al. (1999). We can no longer assume that correlations among common and unique factors are zero, nor that correlations among unique factors are zero, due to sampling variability. Instead, we represent these covariances explicitly as $\mathbf{C}_{uu_i}$ for the sample covariance matrix of unique factors, $\mathbf{C}_{cu_i}$ for the sample covariance matrix of common and unique factors, and $\mathbf{C}_{uc_i}$ for the sample covariance matrix of unique and common factors (MacCallum et al., 1999). For example, even if items are unidimensional and congeneric in the population, they may still have small error covariances or small shared secondary loadings in the sample, which would inflate $\mathbf{C}_{uu_i}$ if the population-generating model was fit (Bandalos, 2002). The item-level sample covariance structure is then given by

$$\mathbf{C}_i = \mathbf{\Lambda}_i\mathbf{C}_{cc_i}\mathbf{\Lambda}_i' + \mathbf{\Lambda}_i\mathbf{C}_{cu_i}\mathbf{\Psi}_i' + \mathbf{\Psi}_i\mathbf{C}_{uc_i}\mathbf{\Lambda}_i' + \mathbf{\Psi}_i\mathbf{C}_{uu_i}\mathbf{\Psi}_i', \tag{6}$$

[4] If we had summed rather than averaged items to form parcels, Equations (4) and (5) imply that unstandardized parcel loadings would be the sum of constituent unstandardized item loadings in the population.

[5] Suppose we were instead comparing *standardized* item versus parcel solutions in the population (i.e., item variances, parcel variances, and factor variances all 1.0). This decrease in residual variance of parcel-indicators as items/parcel increase would force standardized parcel loadings to systematically increase with items/parcel to ensure that parcel variances stay at 1.0.

[6] If we had summed items to form parcels, Equations (4) and (5) imply that residual variance of parcel-indicators would be the sum of the residual variances of constituent item-indicators.

where $\mathbf{C}_{cc_i}$ is the sample covariance matrix of common factors. This means that we can denote the parcel-level sample covariance structure by

$$\mathbf{C}_p = \mathbf{A}(\boldsymbol{\Lambda}_i\mathbf{C}_{cc_i}\boldsymbol{\Lambda}_i')\mathbf{A}' + \mathbf{A}(\boldsymbol{\Lambda}_i\mathbf{C}_{cu_i}\boldsymbol{\Psi}_i')\mathbf{A}' + \mathbf{A}(\boldsymbol{\Psi}_i\mathbf{C}_{uc_i}\boldsymbol{\Lambda}_i')\mathbf{A}' + \mathbf{A}(\boldsymbol{\Psi}_i\mathbf{C}_{uu_i}\boldsymbol{\Psi}_i')\mathbf{A}'. \tag{7}$$

And, rewriting $\boldsymbol{\Lambda}_p = \mathbf{A}\boldsymbol{\Lambda}_i$ and $\boldsymbol{\Psi}_p = \boldsymbol{\Psi}_i'\mathbf{A}' = \mathbf{A}\boldsymbol{\Psi}_i$ we have

$$\mathbf{C}_p = \boldsymbol{\Lambda}_p\mathbf{C}_{cc_i}\boldsymbol{\Lambda}_p' + \boldsymbol{\Lambda}_p\mathbf{C}_{cu_i}\boldsymbol{\Psi}_p + \boldsymbol{\Psi}_p\mathbf{C}_{uc_i}\boldsymbol{\Lambda}_p' + \boldsymbol{\Psi}_p\mathbf{C}_{uu_i}\boldsymbol{\Psi}_p. \tag{8}$$

Following MacCallum et al. (2001), it is possible to simplify the expression for the item-level sample covariance structure by representing all lack of fit due to sampling error with $\boldsymbol{\Delta}_{SE_i}$:

$$\mathbf{C}_i = \boldsymbol{\Lambda}_i\mathbf{C}_{cc_i}\boldsymbol{\Lambda}_i' + \boldsymbol{\Psi}_i^2 + \boldsymbol{\Delta}_{SE_i}. \tag{9}$$

This representation means that we can denote the parcel-level sample covariance structure by

$$\mathbf{C}_p = \mathbf{A}\boldsymbol{\Lambda}_i\mathbf{C}_{cc_i}\boldsymbol{\Lambda}_i'\mathbf{A}' + \mathbf{A}\boldsymbol{\Psi}_i^2\mathbf{A}' + \mathbf{A}\boldsymbol{\Delta}_{SE_i}\mathbf{A}'. \tag{10}$$

And, rewriting $\boldsymbol{\Lambda}_p = \mathbf{A}\boldsymbol{\Lambda}_i$ and $\boldsymbol{\Psi}_p^2 = \mathbf{A}\boldsymbol{\Psi}_i^2\mathbf{A}'$ and $\boldsymbol{\Delta}_{SE_p} = \mathbf{A}\boldsymbol{\Delta}_{SE_i}\mathbf{A}'$ we have

$$\mathbf{C}_p = \boldsymbol{\Lambda}_p\mathbf{C}_{cc_i}\boldsymbol{\Lambda}_p' + \boldsymbol{\Psi}_p^2 + \boldsymbol{\Delta}_{SE_p}. \tag{11}$$

With this theoretical background, several points can now be made about the effects of alternative item-to-parcel allocations (i.e., alternative $\mathbf{A}$ matrices) from random or purposive parceling methods, *in the sample*, even when items are unidimensional and congeneric and the item-level model holds exactly in the population.

Structural Parameter Estimates

Upon first observation, the sample covariance matrix of common factors would seem to be invariant across alternate parcel-allocations because the term $\mathbf{C}_{cc_i}$ appears in the parcel-level sample covariance structure Equation (11) for any $\mathbf{A}$. Furthermore, upon first observation, the sample covariance matrix of common factors would seem to be invariant across parcel-level versus item-level solutions because the term $\mathbf{C}_{cc_i}$ appears in both Equations (9) and (11). However, neither will necessarily be the case because the sampling error in the parcel-analysis will be differentially altered for each item-to-parcel allocation matrix $\mathbf{A}$ (that is, $\boldsymbol{\Delta}_{SE_p} = \mathbf{A}\boldsymbol{\Delta}_{SE_i}\mathbf{A}'$)—even though the number of parcels/factor and items/parcel is held fixed. And, "sources of sampling error ... will have a general and random influence on parameter estimates" (MacCallum & Tucker, 1991, p. 507). Moreover, the effects of alternate $\mathbf{A}$ matrices (i.e., effects of parcel-allocation variability) on parameter estimates will be minimized when the effects of sampling error on parameter estimates is minimized. Following MacCallum and Tucker, this will occur when (a) sample size is large; (b) item communalities are high (i.e., unique loadings $\mathbf{\Psi}_i$ are low, so that $\mathbf{C}_{uu_i}$, $\mathbf{C}_{cu_i}$, and $\mathbf{C}_{uc_i}$ matrices are given little weight in Equation (7)); and (c) $\mathbf{C}_{uu_i}$, $\mathbf{C}_{cu_i}$, and $\mathbf{C}_{uc_i}$ matrices have smaller dimensions (e.g., because more items are allocated to each parcel).

Model Fit

It is already well known that parcel-solutions usually have better fit than item-solutions because parceling ameliorates some effects of sampling error on model fit (Bandalos, 2002; Bandalos & Finney, 2001). Our concern here, however, is *not* the differences in model fit between parcel-solutions and item-solutions but rather the variations in model fit across parcel-solutions based on different $\mathbf{A}$ matrices—given that items are unidimensional and congeneric in the population, the item-level model is properly specified, and numbers of items/parcel and parcels/factor are held fixed across allocations. We can see from Equation (10) that, even under these conditions, model fit in the sample will vary across parcel-allocations $\mathbf{A}$ because $\mathbf{\Delta}_{SE_p} = \mathbf{A}\mathbf{\Delta}_{SE_i}\mathbf{A}'$. Note that this result is independent of any distribution assumptions. That is, no distributional assumptions of a particular estimation method need be violated for this result to hold. Therefore, this result should hold even if normally distributed items are parceled and normal theory estimation is used. We can also see from Equation (7) that the influences that decrease the effects of sampling error on model fit will also decrease the variation in model fit across alternative item-to-parcel allocations $\mathbf{A}$. These influences are (a) larger sample size, (b) higher item communalities, and (c) more items per parcel.

Based on the aforementioned theoretical framework, we are now able to propose hypotheses to be tested in the simulation in the next section, which employs random item-to-parcel allocations. We posit that these hypotheses will hold *even* under the following restrictive conditions: items are unidimensional, congeneric, and normally distributed in the population; a properly specified item-generating model is fit to parceled data; and numbers of items/parcel and parcels/factor are fixed across allocations.

Hypothesis 1. As sample size decreases, item communalities decrease, and the item/parcel ratio decreases, structural parameter estimates will be more variable across random item-to-parcel allocations, within-sample.

Hypothesis 2. As sample size decreases, item communalities decrease, and the item/parcel ratio decreases, model fit estimates will be more variable across random item-to-parcel allocations, within-sample.

Hypothesis 3. Although structural parameter estimates and model fit will be affected by random variation due to item-to-parcel allocating and random variation due to sampling, the effects of the former can be minimized. To do so, researchers need to compute and report average structural parameter and model fit estimates *across a distribution of item-to-parcel allocations* for a given sample. We term this distribution a *parcel-allocation distribution.*

Given that Hypotheses 1–2 have been previously investigated across samples within a single allocation (Hau & Marsh, 2004; Marsh et al., 1998; Nasser & Wisenbaker, 2003; Nasser-Abu & Wisenbaker, 2006), but not—as we do here—across allocations within a single sample, it is desirable to relate our present results to such prior results. To this end, when testing these hypotheses we relate the magnitude and consistency of within-sample parcel-allocation variability to the magnitude and consistency of within-allocation sampling variability. We define *within-sample parcel-allocation variability* as the variability in the outcome of interest (parameter estimate or fit measure) across repeated (and, in our simulation, random) assignments of items to parcels for a single sample—and for prespecified, fixed numbers of parcels/factor and items/parcel. Henceforth, we refer to this simply as *parcel-allocation variability* unless the context requires more detail. We define *within-allocation sampling variability* as the variability in the outcome of interest (parameter estimate or fit measure) across random sampling, for a single item-to-parcel allocation. Henceforth, we refer to this simply as *sampling variability* unless the context requires more detail.

SIMULATION METHODS

Simulation Design

Item-level data were generated from a correlated two-factor confirmatory factor analysis (CFA) population model with either 9 or 15 item indicators/factor. Items were normally distributed, congeneric, and unidimensional in the population. These item characteristics were chosen in order to provide proof of concept: that parcel-allocation variability exists and is practically consequential *even* under such restrictive, idealized conditions. When generating item-level data, we manipulated item loading size (four levels) and sample size (six levels). When constructing parcels from items, we varied the number of items per parcel/number of parcels (three levels). This design resulted in a total of $4 \times 6 \times 3 = 72$ cells in our simulation.

The numbers of items/parcel were either 3 items per each of three parcels (using the 9-item generated data), or 3 items per each of five parcels (using the 15-item generated data), or 5 items per each of three parcels (using the 15-item generated data). These item/parcel combinations were chosen so that we could compare the effect of number of items per parcel (three vs. five), for the same number of parcels (three), and so that we could compare the effect of number of parcels (three vs. five), for the same number of items (15).

Sample sizes ($N = 75$, $N = 100$, $N = 125$, $N = 150$, $N = 200$, $N = 250$) were chosen to include those often used in SEM research in psychology with special focus on low sample sizes, where parcel-allocation variability was hypothesized to be greatest. Baumgartner and Hornberg's (1996) review of SEM applications found an average sample size of 178 and MacCallum and Austin (2000) and Hulland, Chow, and Lam (1996) found that 18–22% of studies used samples < 100. Population-generating parameters, and the scale reliabilities they imply, are shown in Table 1. Values of "high" factor loadings (.70) were chosen, using the Spearman-Brown prophecy formula, to imply a scale reliability for the factor $\geq .90$, which is often considered excellent scale reliability (Nunnally & Bernstein, 1994). Values of "medium-high" factor loadings (.60) were chosen to imply a scale reliability for the factor $\geq .80$, which is often considered good scale reliability (Nunnally & Bernstein, 1994). Values for "low" factor loadings (.40) were chosen to imply a scale reliability for the factor that fell below .80. Additionally, a "medium-mixed" loading condition included a range of loadings sizes from medium-high to low, with an average of .50. Population error variances were chosen to make all item variances = 1.0.

TABLE 1
Population-Generating Parameters

Factor (Co-) Variances	*Factor Loadings (Error Variances in Parentheses)*
$\phi_{11} = 1$ $\phi_{22} = 1$ $\phi_{12} = .25$	High loadings: all $\lambda_i = .70$ (all $\psi_i^2 = .51$) implied scale reliability = .94 for 15 items/factor and .90 for 9 items/factor
	Medium-high loadings: all $\lambda_i = .60$ (all $\psi_i^2 = .64$) implied scale reliability = .89 for 15 items/factor and .84 for 9 items/factor
	Medium-mixed loadings: $\lambda_i = .40, .50, .60$, alternating ($\psi_i^2 = .84, .75$, or $.64$) implied scale reliability = .83 for 15 items/factor and .75 for 9 items/factor
	Low loadings: all $\lambda_i = .40$ (all $\psi_i^2 = .84$) implied scale reliability = .74 for 15 items/factor and .63 for 9 items/factor

Sample and Parcel-Allocation Generation

Data were generated so as to allow for the study of two sources of variability within and across design cells: sampling variability and parcel-allocation variability. Consider a single example design cell. Figure 1 shows that, within this cell, 100 samples were independently generated. Additionally, within this cell, 100 random item-to-parcel allocations were independently generated—by randomly assigning the k items per factor for a given sample to p parcels per factor, where k, p, and numbers of items/parcel were fixed across allocations within a design cell. Then, parcel scores were computed by averaging these randomly assigned items. The same set of 100 allocations within that particular cell was saved and used to generate parcels from items in each of the 100 samples in that cell, as shown in Figure 1. Thus samples and parcel-allocations were fully crossed within a cell. Therefore, in a cell, a statistic has a distribution across parcel-allocations within each sample (i.e., a parcel-allocation distribution) and a distribution across samples within parcel-allocation (i.e., a sampling distribution). The full crossing of the 100 samples and 100 allocations in a cell yielded 10,000 parcel-level data sets in that cell. Across all 72 cells, there was a total of 720,000 parcel-level data sets.

For each of these 720,000 parcel-level data sets, a two-factor CFA model with correlated factors and unidimensional, congeneric items was fit using M*plus* 5.2 (L. K. Muthén & Muthén, 1998–2008) and maximum likelihood estimation. Analyses were performed in a covariance metric (i.e., unstandardized parcel solutions were obtained). In each case, the item-level model was true in the population, and the parcel-level model was fit in the sample. Structural parameter estimates (factor correlation) and model fit statistics (e.g., Root Mean Square Error of Approximation (RMSEA), Standardized Root Mean Square Residual (SRMR), Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), χ^2) were

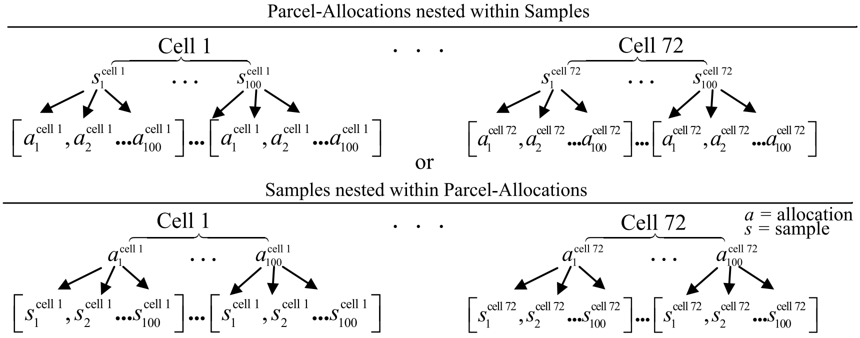

Parcel-Allocations nested within Samples
Cell 1
. . .
Cell 72
$s_1^{\text{cell 1}}$ . . . $s_{100}^{\text{cell 1}}$
$[a_1^{\text{cell 1}}, a_2^{\text{cell 1}} \ldots a_{100}^{\text{cell 1}}] \ldots [a_1^{\text{cell 1}}, a_2^{\text{cell 1}} \ldots a_{100}^{\text{cell 1}}]$
$s_1^{\text{cell 72}}$ . . . $s_{100}^{\text{cell 72}}$
$[a_1^{\text{cell 72}}, a_2^{\text{cell 72}} \ldots a_{100}^{\text{cell 72}}] \ldots [a_1^{\text{cell 72}}, a_2^{\text{cell 72}} \ldots a_{100}^{\text{cell 72}}]$
or
Samples nested within Parcel-Allocations
Cell 1
. . .
Cell 72
a = allocation
s = sample
$a_1^{\text{cell 1}}$ . . . $a_{100}^{\text{cell 1}}$
$[s_1^{\text{cell 1}}, s_2^{\text{cell 1}} \ldots s_{100}^{\text{cell 1}}] \ldots [s_1^{\text{cell 1}}, s_2^{\text{cell 1}} \ldots s_{100}^{\text{cell 1}}]$
$a_1^{\text{cell 72}}$ . . . $a_{100}^{\text{cell 72}}$
$[s_1^{\text{cell 72}}, s_2^{\text{cell 72}} \ldots s_{100}^{\text{cell 72}}] \ldots [s_1^{\text{cell 72}}, s_2^{\text{cell 72}} \ldots s_{100}^{\text{cell 72}}]$

FIGURE 1 Crossed data generation schematic.

recorded for each fitted model. Data analysis involved descriptive comparisons of characteristics (mean, standard deviation, and range) of the parcel-allocation distributions versus sampling distributions of parameter estimates and model fit statistics, for each design cell.

SIMULATION RESULTS

Nonconvergence and Improper Solutions

On average, the proportion of converged and proper allocations per cell was 9,976/10,000 (range 9,302 to 10,000). Proper solutions have positive definite factor and error covariance matrices. Within cell, the average proportion of converged and proper allocations per sample was 99.68/100 (range 46 to 100). Only 36 out of 7,200 samples had $< 90\%$ converged and positive definite allocations. Analyses were rerun including versus omitting improper solutions, and the same pattern of results was obtained. Results presented here omit improper solutions.

Structural Parameter Estimates

When are substantive conclusions about structural parameters sensitive to the particular allocation chosen from the within-sample allocation distribution?

To orient ourselves, we first simply consider the distribution of the factor correlation estimates, $\hat{\phi}$, across 100 parcel allocations within a *single* sample/cell. Recall from Table 1 that in the population $\phi = .25$. Figure 2 presents the distribution of $\hat{\phi}$ across 100 allocations of three items/three parcels for one sample at each loading size and at each sample size. Error bars denote the maximum and minimum $\hat{\phi}$ across allocations within sample. Figure 2 indicates that, given a *single* sample, and, say, low loadings, a researcher could anticipate obtaining $\hat{\phi}$ as low as .13 or as high as .43 (i.e., a range of .30), at $N = 75$, or obtaining $\hat{\phi}$ as low as .01 or as high as .18 (i.e., a range of .17), at $N = 250$—*based simply on the item-to-parcel allocation chosen via random allocating.* For medium-mixed loadings, the range is .30 for $N = 75$ and .10 for $N = 250$ in Figure 2. For medium-high loadings, the range is .18 for $N = 75$ and .05 for $N = 250$. For high loadings, the range is .10 for $N = 75$ and near-zero at $N = 250$.

It is next relevant to consider whether this magnitude of parcel-allocation variability in Figure 2 is "large enough" for a researcher with one sample to actually change substantive conclusions about ϕ. If the result of a researcher's hypothesis test $H_0 : \phi = 0$ was completely robust to choice of parcel-allocation,

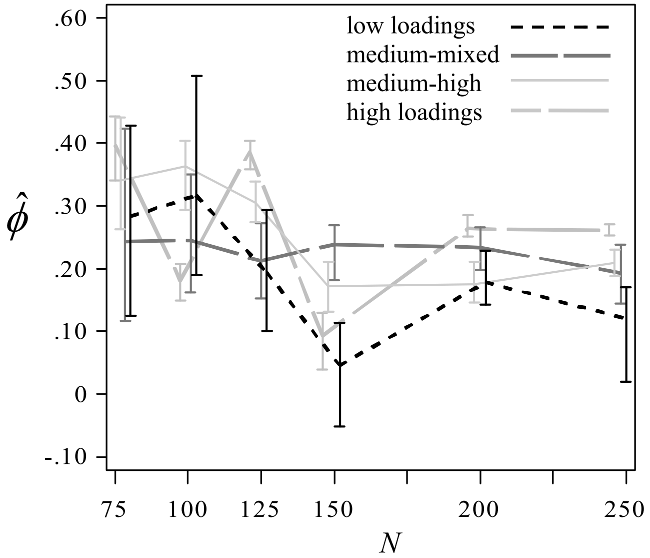


FIGURE 2 The parcel-allocation distribution of a factor correlation estimate ($\hat{\phi}$) within a single sample per cell. *Note*. Each allocation in this figure involved three parcels, each with three items. A parcel allocation distribution of $\hat{\phi}$ is the distribution of this parameter estimate across alternate allocations of items to parcels within a sample.

then all allocations would give a unanimous hypothesis testing result (either all significant or all nonsignificant). If a researcher's hypothesis testing result was maximally sensitive to the allocation chosen, then 50% of allocations would yield statistical significance, and 50% would yield nonsignificance. For the sample with $N = 100$ and low loadings, 64% of allocations gave a statistically significant $\hat{\phi}$ and 36% of allocations gave a nonsigificant $\hat{\phi}$. Similarly, for the sample with $N = 125$ and medium-mixed loadings, 46% of allocations gave a statistically significant $\hat{\phi}$ and 54% of allocations gave a nonsignificant $\hat{\phi}$. But, for any sample in Figure 2 with $N \geq 150$ or medium-high/high loadings, 91–100% of allocations yielded the same significance value. These results show that a researcher's hypothesis test results can sometimes change markedly depending on the particular parcel allocation chosen for their single sample. But a researcher's conclusions about the significance of ϕ are less likely to change across allocations at higher sample size and loading size.

Next, rather than just consider a single sample's $\hat{\phi}$ allocation distribution per cell, we consider 100 samples' $\hat{\phi}$ allocation distributions per cell. Specifically, for the 100 samples per cell, we calculate the range of each sample's $\hat{\phi}$ allocation distribution (i.e., $R^a_{\hat{\phi}}$) and standard deviation of each sample's $\hat{\phi}$ allocation distribution (i.e., $SD^a_{\hat{\phi}}$). Moreover, for each of these 100 samples per cell, we also calculate the proportion of allocations with statistically significant $\hat{\phi}$. This results in 100 $R^a_{\hat{\phi}}$, 100 $SD^a_{\hat{\phi}}$, and 100 proportions of significant $\hat{\phi}$ per cell. For each cell, we plotted the distribution of the 100 $R^a_{\hat{\phi}}$, in Figure 3, Panel 1, and the distribution of the 100 $SD^a_{\hat{\phi}}$, in Figure 4, Panel 1. Due to space constraints, results for high loadings are not presented in Figures 3 and 4; they showed, on average, near-zero allocation variability in $\hat{\phi}$.

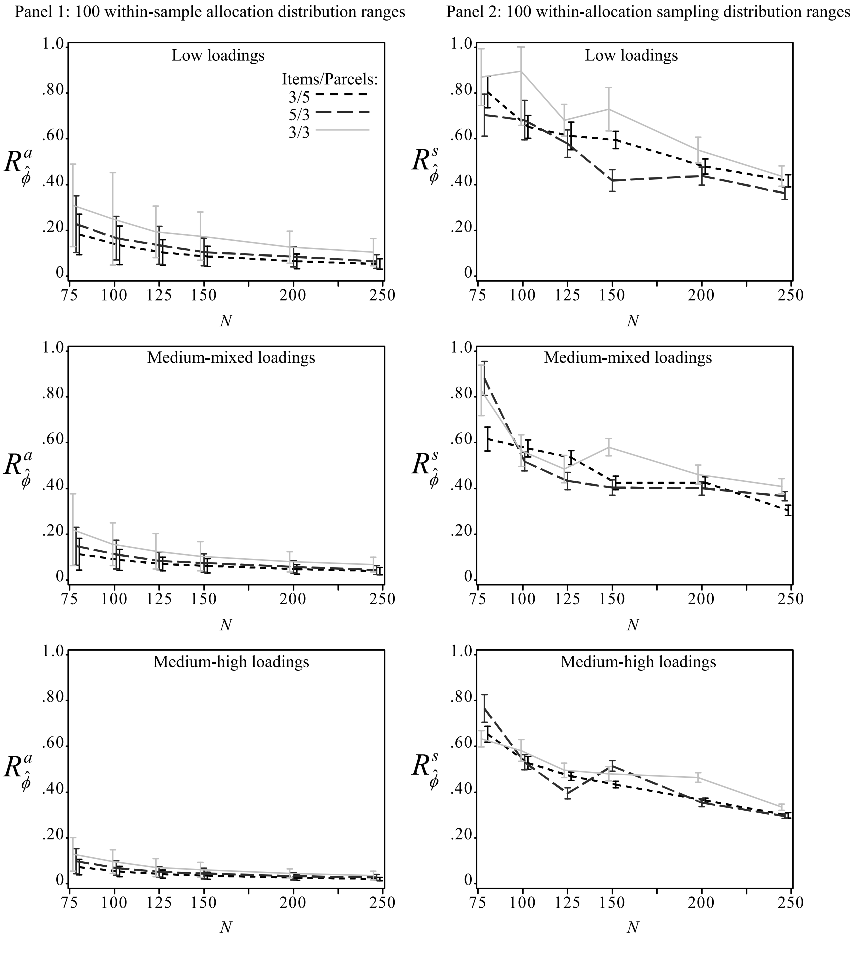
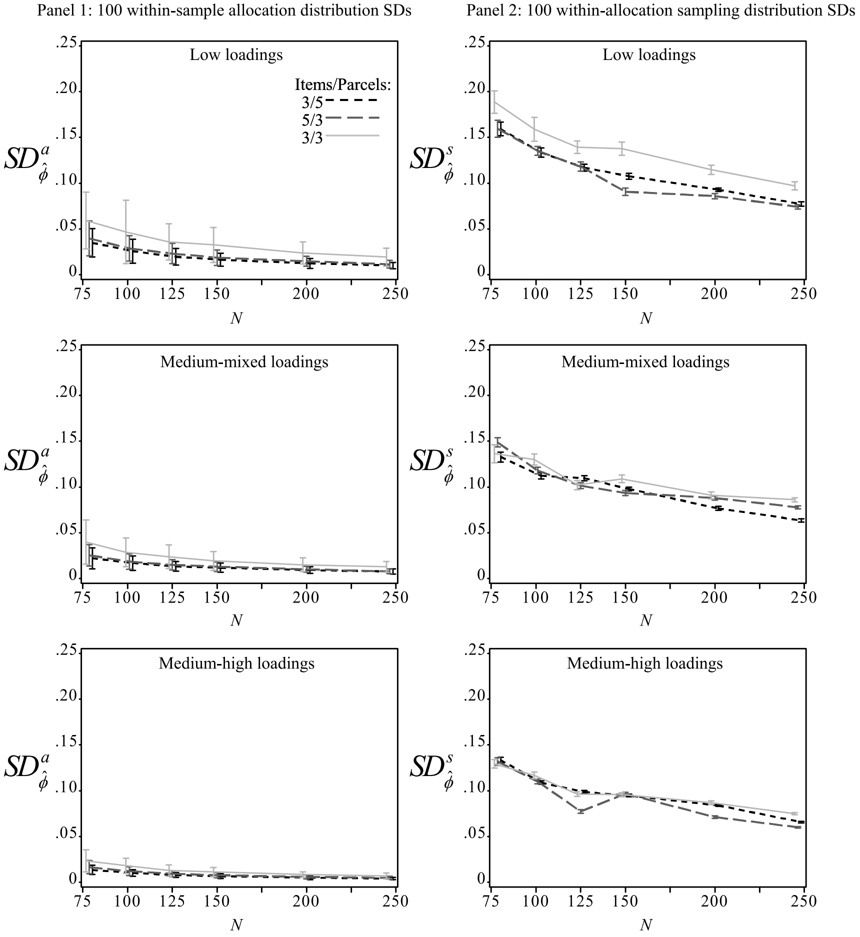


FIGURE 3 Ranges of within-sample $\hat{\phi}$ allocation distributions versus ranges of within-allocation $\hat{\phi}$ sampling distributions.


Panel 1: 100 within-sample allocation distribution SDs

Panel 2: 100 within-allocation sampling distribution SDs

Low loadings

Medium-mixed loadings

Medium-high loadings

FIGURE 4 Standard deviations of within-sample $\hat{\phi}$ allocation distributions versus standard deviations of within-allocation $\hat{\phi}$ sampling distributions.

In Figure 3, Panel 1, error bars denote ±2 *SD* around the cell mean of the $R^a_{\hat{\phi}}$'s. The cell mean of $R^a_{\hat{\phi}}$'s occurs where the line intersects the error bars. For few parcels and few items/parcel, the average $R^a_{\hat{\phi}}$ is sizable (e.g., $\geq .10$ correlation units) at low loadings when $N \leq 250$, at medium-mixed loadings when $N \leq 150$, and at medium-high loadings when $N \leq 100$. This means that the amount of allocation variability depicted for single samples in Figure 2 is typical of the other samples in those cells. For few parcels and many items/parcel, the average $R^a_{\hat{\phi}}$ is sizable at low loadings when $N \leq 150$, for medium-mixed loadings when $N \leq 100$, and for medium-high loadings when $N = 75$. For many parcels, average $R^a_{\hat{\phi}}$ is sizable at low loadings, when $N \leq 125$, and at medium-mixed loadings, when $N = 75$. That is—given unidimensional and congeneric items—only for medium loadings together with larger N is the average $R^a_{\hat{\phi}}$ near-zero for any item/parcel combination.

A similar pattern occurs in Figure 4, Panel 1, as was found in Figure 3, Panel 1. In Figure 4, Panel 1, error bars now denote ±2 *SD* around the cell mean of $SD^a_{\hat{\phi}}$'s and the cell mean of the $SD^a_{\hat{\phi}}$'s occurs where the line intersects the error bars. On average, the within-sample allocation distribution standard deviations $SD^a_{\hat{\phi}}$ are largest for $N = 75$ and low loadings—specifically, .04–.06 correlation units in magnitude. Average $SD^a_{\hat{\phi}}$ decreases for larger loadings and for larger N. If larger loadings are combined with larger N, average $SD^a_{\hat{\phi}}$ is near zero—again, assuming unidimensional, congeneric normal items and no model error.

Finally, Table 2 shows the degree to which substantive conclusions about structural parameter estimates are sensitive to parcel-allocation variability. Specifically, Table 2 shows the proportion of samples per cell in which > 5% of allocations *per sample* switch statistical significance of $\hat{\phi}$ (i.e., nonsignificant to/from significant) across that sample's allocation distribution. Samples with $N \leq 150$ and medium-mixed or low loadings most commonly show such switching of structural parameter estimate significance level across allocations within sample. However, nearly a quarter of allocations per sample switch statistical significance

TABLE 2
Percentage of Samples in Which > 5% of Allocations/Sample Changed $\hat{\phi}$ Significance Level

		Sample Size					
Loading	*Items/Parcels*	*75*	*100*	*125*	*150*	*200*	*250*
Low	3/5	18	23	17	15	12	7
Low	5/3	31	27	29	22	16	11
Low	3/3	43	44	38	30	25	20
Medium-mixed	3/5	26	25	20	11	8	4
Medium-mixed	5/3	23	20	17	13	7	12
Medium-mixed	3/3	42	27	19	23	14	11
Medium-high	3/5	14	12	5	2	1	3
Medium-high	5/3	13	12	8	9	2	3
Medium-high	3/3	24	23	19	13	11	7
High	3/5	11	9	4	2	3	1
High	5/3	12	7	9	4	3	0
High	3/3	15	7	9	6	6	4

even for (a) low loadings + high samples or (b) medium-high loadings + low samples—in the context of few items/parcel and few parcels.

Does the variation in structural parameter estimates differ across-allocations within-sample versus across-samples within allocation?

To answer this question, we reanalyzed the simulation data in the opposite way. That is, instead of looking at within-sample allocation variability in structural parameter estimates, we looked at within-allocation sampling variability in structural parameter estimates. Specifically, for the 100 allocations per cell, we calculated the range ($R^s_{\hat{\phi}}$) and standard deviation ($SD^s_{\hat{\phi}}$) of each allocation's $\hat{\phi}$ sampling distribution. This resulted in 100 $R^s_{\hat{\phi}}$'s and 100 $SD^s_{\hat{\phi}}$'s per cell. For each cell, we plotted the distribution of the 100 $R^s_{\hat{\phi}}$, in Figure 3, Panel 2, and the distribution of the 100 $SD^s_{\hat{\phi}}$, in Figure 4, Panel 2.

In Figure 3, Panel 2, error bars denote ±2 *SD* around the cell mean of $R^s_{\hat{\phi}}$'s and the cell mean of the $R^s_{\hat{\phi}}$'s occurs where the line intersects the error bars. Figure 3, Panel 2, shows that the average within-allocation sampling distribution ranges, $R^s_{\hat{\phi}}$, are larger than the average $R^a_{\hat{\phi}}$, and unlike parcel-allocation variability, sampling variability does not decrease all the way to zero for larger loading and sample sizes. In Figure 4, Panel 2, error bars denote ±2 *SD* around the cell mean of $SD^s_{\hat{\phi}}$'s and the cell mean of the $SD^s_{\hat{\phi}}$'s occurs where the line intersects the error bars.

Comparing Figure 4, Panel 1, with Panel 2, average within-allocation sampling distribution standard deviations, $SD^s_{\hat{\phi}}$, are always greater than the average $SD^a_{\hat{\phi}}$, but the variability (error bars) in $SD^a_{\hat{\phi}}$ can be greater than the variability (error bars) in $SD^s_{\hat{\phi}}$, particularly for few items/parcel and few parcels. Moreover, as in Figure 3, Panels 1 and 2, both parcel-allocation variability and sampling variability are maximized under the same data conditions: lower loadings, lower sample size, and fewer items/parcels.

After controlling for parcel-allocation error, does sampling variability in structural parameter estimates remain and, after controlling for sampling error, does allocation variability in structural parameter estimates remain?

To answer this question, we calculate the average $\hat{\phi}$, across allocations within-sample (i.e., $M^a_{\hat{\phi}}$). Hence, there are 100 $M^a_{\hat{\phi}}$'s per cell. And we calculate the average $\hat{\phi}$, across-samples within-allocation (i.e., $M^s_{\hat{\phi}}$). Hence, there are 100 $M^s_{\hat{\phi}}$'s per cell. Figure 5, Panel 1, tells us that if we ameliorate/eliminate parcel-allocation error by averaging across 100 allocations within each sample, how

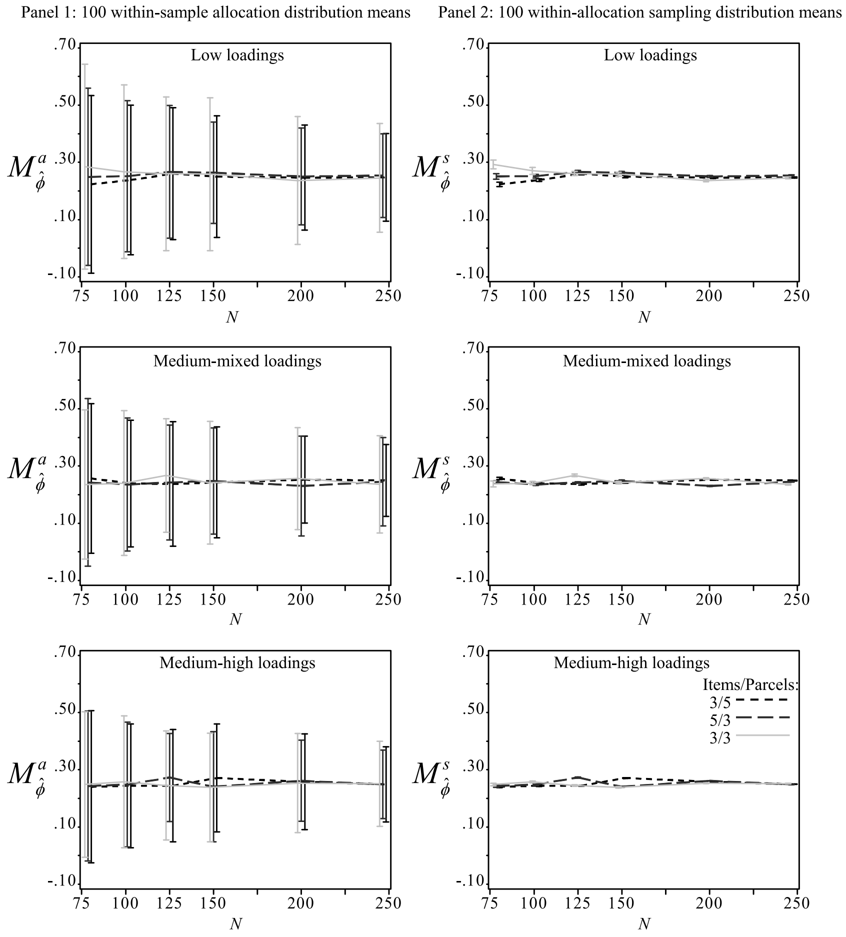


FIGURE 5 Means of within-sample $\hat{\phi}$ allocation distributions versus means of within-allocation $\hat{\phi}$ sampling distributions.

much sample-to-sample variability will remain in the distribution of $M^a_{\hat{\phi}}$'s per cell, and will the cell mean of $M^a_{\hat{\phi}}$'s be on target at the population-generating parameter? Figure 5, Panel 2, tells us that if we ameliorate/eliminate sampling error by averaging across 100 samples within each allocation, how much allocation-to-allocation variability will remain in the distribution of $M^s_{\hat{\phi}}$'s per cell, and will the cell mean of the $M^s_{\hat{\phi}}$ be on target at the population-generating parameter? Note that, in Figure 5, Panel 1, error bars now denote ±2 *SD* around the cell mean of the $M_{\hat{\phi}}^{a}$'s and in Figure 5, Panel 2, error bars now denote ±2 *SD* around the cell mean of the $M_{\hat{\phi}}^{s}$'s. In Figure 5, Panel 1, we can see from where the line intersects the error bars that, after averaging across allocations within-sample, the mean of the distribution $M_{\hat{\phi}}^{a}$ of closely matches the population-generating value for all cells. Likewise, in Figure 5, Panel 2, we can see from where the line intersects the error bars that, after averaging across samples within-allocation, the mean of the distribution of $M_{\hat{\phi}}^{s}$ matches the population-generating value for all cells. Importantly, in Figure 5, Panel 1, after averaging across parcel-allocations within-sample, *considerable sample-to-sample variability remains* among the 100 $M_{\hat{\phi}}^{a}$ per cell. Whereas, in Figure 5, Panel 2, after averaging across samples within-allocation, essentially *no allocation-to-allocation variability remains* in the 100 $M_{\hat{\phi}}^{s}$ per cell. (Of course, within any *single* sample, across-allocation variability similar to that shown in Figure 2 would still remain.) We discuss in a later section how this point demonstrated in Figure 5 motivates our solution to the problem of parcel-allocation variability in finite samples with unreliable items. To summarize this point, there is little parcel-allocation variability in the absence of sampling error, but there can be substantial sampling variability in the absence of parcel-allocation variability.

Model Fit

When are substantive conclusions about model adequacy sensitive to the particular allocation chosen from the within-sample allocation distribution?

To orient ourselves, we first simply consider the distribution of model fit indices across the 100 parcel allocations within each *single* sample per cell from Figure 2. Recall that this set of single samples from Figure 2 included one sample at each loading size and at each N, for three items/three parcels. Table 3 shows that parcel-allocation variability in model fit indices within this set of single samples exists at all sample sizes and all loading sizes. Moreover, the ranges of these within-sample allocation distributions often—even at some high loading or high sample sizes—include conventional cutoff values of CFI, TLI, RMSEA, and SRMR. In such cases, substantive conclusions about model adequacy could change based on the particular allocation chosen. Also, the ranges of χ^2 across allocation per sample are around twice the mean χ^2 statistic per sample.

Next, we consider whether the parcel-allocation distributions of model fit indices, which were documented for single samples in Table 3, are representative of all 100 samples per cell and all cells. To consolidate results and make their implications for practice maximally clear, we made the following three simplifications. First, because the same general pattern of results was found for

TABLE 3
Distribution of Model Fit Statistics Across Allocations Within a Single Sample/Cell for Three Items per Three Parcels

		TLI			*SRMR*			*CFI*			*RMSEA*			*Chi-Square (df = 8)*		
Loadings	*N*	*M*	*SD*	*Range*	*M*	*SD*	*Range*	*M*	*SD*	*Range*	*M*	*SD*	*Range*	*M*	*SD*	*Range*
Low	75	.998	.114	.514	.053	.013	.057	.975	.037	.151	.034	.039	.129	8.36	3.51	15.52
	100	.920	.114	.584	.054	.015	.066	.950	.052	.229	.058	.045	.163	11.72	5.20	26.06
	125	.984	.078	.405	.043	.012	.048	.978	.029	.151	.034	.037	.136	9.16	4.64	24.66
	150	.996	.047	.235	.035	.007	.033	.989	.017	.079	.023	.027	.095	8.35	3.53	16.58
	200	1.027	.040	.232	.026	.005	.025	.997	.011	.069	.007	.016	.076	6.18	2.84	15.45
	250	.995	.050	.207	.031	.007	.033	.988	.019	.070	.019	.023	.079	8.53	3.85	17.93
Medium-mixed	75	.947	.071	.339	.058	.015	.065	.968	.032	.136	.064	.047	.161	11.26	4.33	21.34
	100	.987	.048	.231	.044	.011	.052	.986	.018	.070	.037	.037	.116	9.11	3.90	17.82
	125	1.009	.035	.160	.031	.008	.035	.995	.010	.050	.018	.028	.100	7.17	3.59	16.64
	150	.993	.029	.147	.035	.008	.036	.992	.011	.052	.027	.029	.100	8.92	3.60	17.38
	200	1.010	.021	.100	.027	.008	.035	.997	.005	.027	.011	.019	.071	6.51	3.17	15.11
	250	.994	.023	.108	.030	.009	.041	.993	.008	.039	.025	.025	.085	9.26	4.43	21.18
Medium-high	75	.944	.066	.319	.059	.014	.064	.967	.032	.133	.073	.050	.186	12.35	5.23	26.05
	100	1.006	.022	.125	.033	.008	.052	.997	.007	.040	.018	.028	.118	7.21	3.03	18.21
	125	1.009	.022	.108	.030	.008	.035	.997	.006	.029	.014	.025	.094	6.74	3.18	15.96
	150	.994	.031	.139	.031	.010	.043	.992	.012	.050	.030	.034	.117	9.02	5.02	23.57
	200	.996	.014	.068	.028	.007	.030	.996	.005	.022	.025	.026	.083	9.02	3.79	17.19
	250	.992	.016	.078	.025	.007	.027	.994	.007	.031	.030	.027	.094	10.39	4.77	23.80
High	75	.995	.024	.128	.034	.011	.045	.994	.009	.049	.040	.044	.161	8.96	4.14	21.82
	100	1.010	.015	.076	.028	.007	.030	.998	.004	.020	.011	.023	.091	6.15	2.78	12.84
	125	.991	.019	.096	.031	.010	.043	.993	.008	.037	.043	.040	.138	10.44	5.09	26.27
	150	.998	.015	.088	.027	.008	.037	.997	.006	.035	.023	.031	.127	8.49	4.24	25.61
	200	.994	.012	.059	.027	.009	.040	.996	.005	.024	.033	.031	.109	10.56	4.98	25.04
	250	.998	.007	.030	.021	.005	.023	.998	.003	.010	.021	.022	.069	8.88	3.46	15.39

Notes. TLI = Tucker-Lewis Index; SRMR = Standardized Root Mean Square Residual; CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation.

RMSEA, SRMR, CFI, TLI, and χ^2, we just focus on RMSEA. Second, because in practice fit in SEM is commonly evaluated using extremity cutoff criteria, we focus on the parcel-allocation distribution of RMSEA *nonclose fit* for all 100 samples per cell and for all cells. We selected a recommended cutoff value for RMSEA nonclose fit ($\geq .06$; Hu & Bentler, 1999) that yielded 21–31% of converged, proper solutions per cell nonclose fitting at $N = 75$; 12–26% at $N = 100$, 6–20% at $N = 125$, 3–17% at $N = 150$, 1–10% at $N = 200$, and <1–7% at $N = 250$. (Because cutoff values are ultimately arbitrary, we performed a sensitivity analysis using a variety of cutoff values for RMSEA (.05, .06, .08), which produced the same overall pattern of results). As a third simplification, we present results only for the low and high loading conditions, not the medium-mixed or medium-high loading conditions, because results were similar across all loading sizes.

Figure 6 addresses how the parcel-allocation distribution of RMSEA nonclose fit varies from sample to sample within cell. Figure 6 contains 36 boxplots, each corresponding to a cell of the simulation design (i.e., 72 total cells—36 cells for the medium loading results not shown). The data points for constructing each cell's boxplot are percentages of nonclose fitting (but converged and proper) allocations per sample. Thus if model fit was completely robust to parcel-allocation variability, we would expect to see that either all allocations per sample were nonclose fitting (i.e., all samples' data points at 100) or no allocations per sample were nonclose fitting (i.e., all samples' data points at 0) in the boxplot. Instead, there are no samples in any cell where 100% of allocations are nonclose fitting and there are only four cells where nearly all

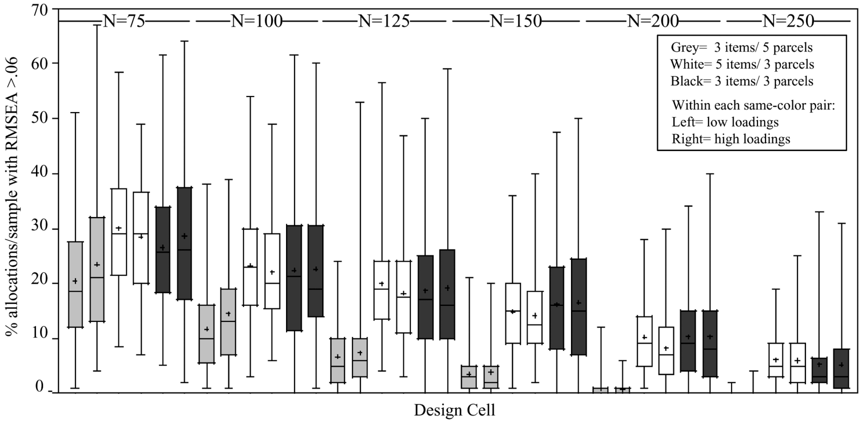


FIGURE 6 Distribution of RMSEA nonclose fit: Parcel-allocations nested in samples. *Note.* RMSEA = Root Mean Square Error of Approximation.

samples had 0% of allocations nonclose fitting. We can glean more detailed information about parcel allocation variability in nonclose fit from inspecting particular characteristics of the boxplots: the + signs, box lengths, and whisker lengths.

The + sign denotes the average percentage of nonclose fitting allocations/sample within that cell. Comparing the location of the +'s across cells, on average, 20–30% of allocations/sample are nonclose fitting at $N = 75$ regardless of loading size. For three parcels/factor, on average, 20–25% of allocations/sample are nonclose fitting at $N = 100$ or 125; 15% of allocations/sample are nonclose fitting at $N = 150$; and $\leq$ 10% of allocations/sample are nonclose fitting at $N = 200$ or 250—regardless of loading size. Whereas, for five parcels/factor, $\leq$ 10% of allocations/sample are nonclose fitting for $N \geq 125$—regardless of loading size. In other words, parcel-allocation variability in model fit (Figure 6) is meaningful under a wider variety of sample size and loading conditions than was parcel-allocation variability in structural parameter estimates (Panel 1 of Figures 3 and 4). Specifically, for structural parameter estimates, medium-high to high loadings and/or sample sizes > 150 generally corresponded with little allocation variability, whereas for model fit, high loadings at $N = 200$ can still lead to frequent (more than 1 out of 10) allocations switching between inadequate and adequate fit statistics within a sample.

The length of each box in Figure 6 denotes the interquartile range of the percentage of allocations/sample that are nonclose fitting within that cell. The length of each set of whiskers in Figure 6 denotes the range of the percentage of allocations/sample that are nonclose fitting within that cell. For example, for high loadings at $N = 125$ and five items per three parcels (Box #16 from left), one quarter of samples have over 24% ill-fitting allocations and one quarter of samples have less than 11% ill-fitting allocations. Together, box length and whisker length for cells in Figure 6 indicate considerable variability from sample to sample in the the percentage of nonclose fitting allocations/sample, regardless of loading size. Reduced variability occurred for many parcels and $N \geq 150$ but only as a by-product of the fact that the net amount of nonclose fit from sampling error approached zero under these conditions.

Is the amount of nonclose-fit per cell differentially distributed within-samples versus within-allocations?

This second question is addressed by comparing features of the boxplots in Figure 6 with Figure 7. Each of the 36 boxplots in Figure 7 correspond to a cell of the simulation design—presented in the same order as in Figure 6. Whereas the data points for making boxplots in Figure 6 were percentages of nonclose fitting allocations per sample, the data points for making boxplots in Figure 7 are percentages of nonclose fitting samples per allocation. Whereas the Figure 6

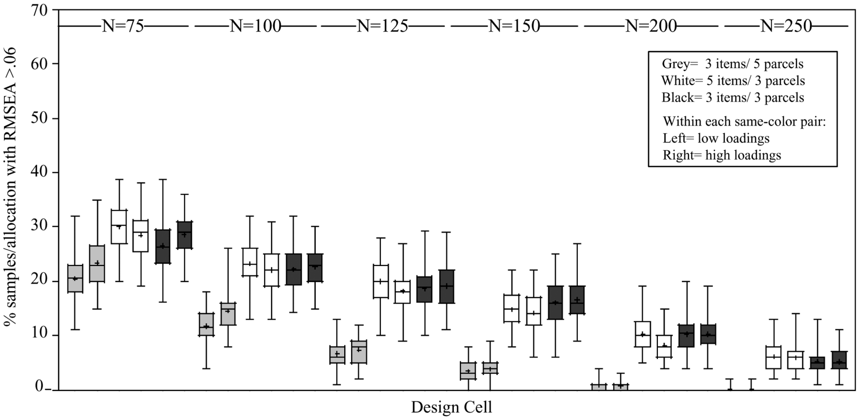


FIGURE 7 Distribution of RMSEA nonclose fit: Samples nested in parcel-allocations. *Note.* RMSEA = Root Mean Square Error of Approximation.

data points can be thought of as cluster sizes where nonclose fitting allocations are Level 1 units and samples are clusters, the Figure 7 data points can be thought of as cluster sizes where nonclose fitting samples are Level 1 units and allocations are clusters. Using these terms, the + sign denotes average cluster size in both panels. In both panels, box length denotes the interquartile range of these cluster sizes, and whisker length denotes the range of these cluster sizes.

The following pattern is evident: (a) average cluster size (+ sign) is often quite similar from Figure 6 to 7, but (b) variability of cluster sizes is typically much greater in Figure 6 than in Figure 7, both in terms of the interquartile range (box length) and range (whisker length). That is, clusters in Figure 7 are more balanced than clusters in Figure 6. To illustrate this pattern of results, consider the boxplot that has $N = 150$, high loadings, and three items per three parcels (Boxplot #24 from left). Its average cluster size in Figure 6 is large: 17% allocations/sample nonclose fitting. Its variability in cluster sizes in Figure 6 is also large (interquartile range of cluster sizes: 8–24% allocations/sample nonclose fitting; full range: 0–49%). Its average cluster size in Figure 7 is similarly large: 17% of samples/allocation nonclose fitting. However, its variability in cluster sizes in Figure 7 is modest (interquartile range of cluster sizes: 14–20% samples/allocation nonclose fitting; full range: 10–27%). This pattern of results means that there is greater predictability and regularity in the sampling variability of RMSEA (Figure 7) than in the parcel-allocation variability of RMSEA (Figure 6). That is, for a given set of data characteristics (N, loading size, numbers of items/parcel and parcels/factor) the likelihood of obtaining nonclose RMSEA fit will be similar from one sample to another, within a particular allocation, but the likelihood of obtaining nonclose RMSEA fit will differ considerably from one allocation to another, within a particular sample.

In sum, the amount of nonclose fit *is* differentially distributed among allocations within samples (much more variable) versus among samples within allocations (much less variable). But this differential distribution of nonclose fit is minimized when variability in cluster sizes is minimized and/or when nonclose fit is minimized altogether—that is, for large N or for medium-to-large N plus larger numbers of parcels.

After controlling for parcel-allocation error, does sample-to-sample variability in model nonclose fit remain and, after controlling for sampling error, does allocation-to-allocation variability in model nonclose fit remain?

Our approach for answering this question, as in the parameter estimates section, is to average over allocations within sample and then consider whether meaningful sample-to-sample variability remains in the distribution of these means (Table 4, Panel 1). We also average over samples within allocation and then consider whether meaningful allocation-to-allocation variability remains in the distribution of these means (Table 4, Panel 2). Table 4, Panel 2, indicates that upon ameliorating/removing sampling error, parcel-allocation variability becomes practically insignificant. That is, the rows of maximum cluster means never exceed .06, and so—controlling for sampling error—parcel-allocation variability would rarely, if ever, lead us to erroneously conclude that a correctly specified model was nonclose fitting. This means that if a researcher were to report average model fit indices across a distribution of parcel-allocations (rather than report model fit indices for a single allocation), the effects of parcel-allocation variability on model fit would be greatly minimized and no longer of practical concern (even for low N, small loadings, and few items/few parcels). On the other hand, Table 4, Panel 1, indicates that upon ameliorating/removing parcel-allocation error, considerable sampling variability is still present; for example, when $N < 150$, the range of cluster means includes .06. Note that the mean of the cluster mean columns are the same, to three decimals, in Panel 1 versus Panel 2.

EMPIRICAL EXAMPLE

Our simulation provided proof of concept of parcel-allocation variability under conditions where it was previously thought not to exist: unidimensional, congeneric, normal items with no model error. However, these are sterile conditions and so our simulation results likely represent a conservative estimate of the amount of parcel-allocation variability that would be encountered in the real

TABLE 4
RMSEA Nonclose Fit Controlling for Parcel-Allocation Error
Vs. Controlling for Sampling Error

		1. Average Across Allocations Within Sample					*2. Average Across Samples Within Allocation*			
Loadings	*Items/Parcels*	*Min of Cluster Means*	*Max of Cluster Means*	*M of Cluster Means*	*SD of Cluster Means*		*Min of Cluster Means*	*Max of Cluster Means*	*M of Cluster Means*	*SD of Cluster Means*
Low	3/5	.006	.053	.029	.010	$N = 75$	.022	.036	.029	.003
	5/3	.013	.070	.036	.012		.027	.046	.036	.004
	3/3	.008	.072	.032	.014		.023	.042	.032	.004
High	3/5	.010	.068	.033	.011		.026	.040	.033	.003
	5/3	.011	.056	.033	.011		.024	.043	.033	.004
	3/3	.003	.073	.034	.016		.025	.041	.034	.003
Low	3/5	.009	.045	.024	.008	$N = 100$	.018	.030	.024	.002
	5/3	.005	.060	.028	.010		.018	.036	.028	.004
	3/3	.002	.071	.027	.014		.019	.035	.027	.003
High	3/5	.007	.048	.027	.009		.021	.034	.027	.003
	5/3	.008	.055	.027	.010		.017	.037	.027	.004
	3/3	.004	.069	.028	.013		.019	.035	.028	.003
Low	3/5	.008	.041	.021	.008	$N = 125$	.014	.026	.021	.002
	5/3	.010	.066	.026	.009		.018	.033	.026	.003
	3/3	.002	.052	.025	.012		.015	.031	.025	.003
High	3/5	.006	.056	.022	.008		.017	.027	.022	.002
	5/3	.006	.054	.025	.009		.019	.032	.025	.003
	3/3	.004	.061	.026	.011		.016	.035	.026	.004
Low	3/5	.004	.044	.019	.007	$N = 150$	.012	.024	.019	.002
	5/3	.007	.043	.023	.008		.017	.032	.023	.003
	3/3	.006	.056	.024	.011		.017	.030	.024	.003
High	3/5	.007	.038	.020	.007		.014	.025	.020	.002
	5/3	.008	.043	.022	.008		.016	.030	.022	.003
	3/3	.002	.053	.024	.012		.017	.034	.024	.003
Low	3/5	.005	.037	.016	.006	$N = 200$	.012	.019	.016	.002
	5/3	.008	.038	.021	.007		.014	.026	.021	.002
	3/3	.004	.046	.020	.009		.016	.026	.020	.002
High	3/5	.003	.031	.016	.006		.010	.020	.016	.002
	5/3	.006	.036	.018	.007		.010	.023	.018	.003
	3/3	.003	.050	.020	.010		.016	.026	.020	.002
Low	3/5	.004	.027	.014	.005	$N = 250$	.010	.017	.014	.001
	5/3	.005	.032	.018	.007		.011	.022	.018	.002
	3/3	.003	.044	.017	.008		.012	.022	.017	.002
High	3/5	.002	.030	.013	.005		.010	.017	.013	.001
	5/3	.003	.035	.017	.007		.011	.021	.017	.002
	3/3	.001	.040	.016	.008		.012	.021	.016	.002

world. To give a more realistic depiction of the magnitude of parcel-allocation variability encountered in the real world, we provide an empirical example. This empirical example involves items from the Big Five personality factors (NEO Personality Inventory; Costa & McCrae, 1985) for which parceling has been previously recommended when the goal is understanding structural relationships rather than dimensionality testing (e.g., Little et al., 2002) and has been extensively used (e.g., Clara, Cox, & Enns, 2003; Benet-Martinez & Karakitapoglu, 2003; Saucier, 2002). Indeed, a brief literature review found over 20 articles applying parceling to Big Five inventory items.

This empirical example also involves a more complex structural model than the simple model considered in the simulation. Following Asendorpf and Wilpers (1998), our example model posits that two correlated Big Five factors (agreeableness and conscientiousness) predict two correlated kinds of social support: belonging (perceived availability of people to do things with) and tangible support (perceived availability of material aid). Figure 8 presents a path diagram of this four-factor model. Prior literature suggested all four factors should each be unidimensional. The eighteen 5-point agreeableness items were allocated 100 times to three parcels of 6 items. The eighteen 5-point conscientiousness items were also allocated 100 times to three parcels of 6 items. The 12 binary belonging items (from the Cohen-Hoberman, 1983, Interpersonal Support Evaluation List) were allocated 100 times to three parcels of 4 items, as were the 12 binary tangible items from the same measure. This parceling of nonmetric items introduces some model error, again making this example more realistic.

Participants were $N = 102$ undergraduate students in the 1988 Computer Assisted Panel Study (Latane, 1989). Note that these empirical example results can be compared most closely with the condition of our simulation with low loadings, $N = 100$, and five items per three parcels. (Although an item-level

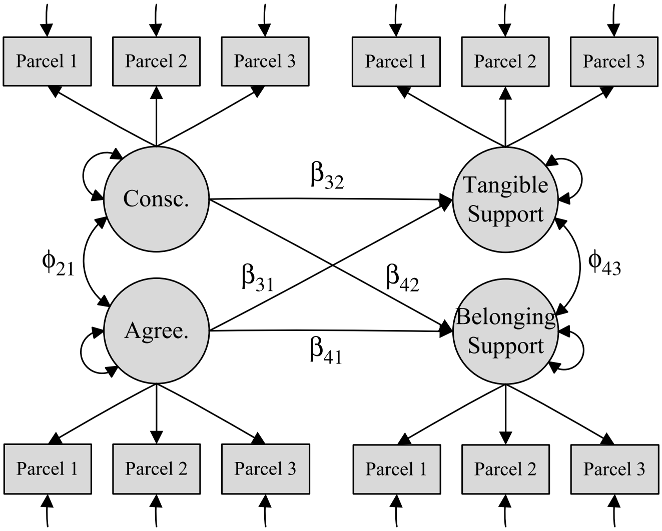


FIGURE 8 Path diagram of empirical example model.

Notes. Consc. = Conscientiousness; Agree. = Agreeableness.

analysis of the empirical data encountered estimation problems—not surprising given the large number of items for an N of 100—the across-allocations average of unstandardized parcel loadings ranged from below to above our simulation's low item loadings conditions: .29–.34 for agreeableness, .15–.16 for conscientious, .45–.48 for tangible support, and .34–.35 for belonging support).

Results in Table 5 show that under these real-world conditions (e.g., more complex model, some model error), allocation variability in structural parameter estimates was larger than in the most nearly comparable simulation cells. For example, the across-allocation ranges of structural parameter estimates were each one to four times the size of their parameter's point estimate. Importantly, the range and standard deviation of allocation distributions for structural parameter estimates were so large that a researcher's hypothesis test results could change

TABLE 5
Within-Sample Parcel-Allocation Variability in an Empirical Example

	Across-Allocation Distribution					
Structural Parameters	*M*	*SD*	*Minimum*	*Maximum*	*Range*	*% Allocations w/ Significant Est.*
$\hat{\beta}_{31}$	−.450	.110	−.756	−.142	.615	78
$\hat{\beta}_{32}$	.277	.099	−.021	.521	.542	11
$\hat{\beta}_{41}$	−.233	.080	−.437	−.005	.432	6
$\hat{\beta}_{42}$	−.076	.058	−.244	.062	.306	0
$\hat{\phi}_{21}$	.256	.067	.079	.407	.328	100
$\hat{\phi}_{43}$	.489	.061	.342	.616	.274	20
$\hat{\beta}_{31}$ *SE*	.205	.033	.148	.341	.193	
$\hat{\beta}_{32}$ *SE*	.191	.024	.148	.265	.118	
$\hat{\beta}_{41}$ *SE*	.186	.024	.143	.295	.152	
$\hat{\beta}_{42}$ *SE*	.178	.013	.144	.224	.080	
$\hat{\phi}_{21}$ *SE*	.162	.021	.136	.226	.089	
$\hat{\phi}_{43}$ *SE*	.143	.013	.119	.191	.072	
Model Fit						
Chi-square (*df* = 48)	62.378	10.609	39.062	87.343	48.281	41
Chi-square *p* value	.162	.190	.000	.818	.818	
CFI	.933	.043	.843	1.000	.157	
TLI	.910	.064	.785	1.059	.274	
RMSEA	.049	.024	.000	.090	.090	
SRMR	.068	.007	.054	.088	.034	

Note. SE = Standard Error; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual.

depending on the particular allocation chosen (see the Table 5 column listing the percentage of allocations yielding significant estimates). For example, the correlation between tangible and belonging support was significant in 20% of allocations and nonsignificant in the other 80%. Moreover, a researcher's substantive conclusions about whether the model fits well also change markedly across allocations in Table 5. For example, the chi-square is significant in 41% of allocations. Also, a one standard deviation span of the allocation distributions of RMSEA, CFI, and TLI include values ranging from close fit to nonclose fit, as does the range of SRMR.

DISCUSSION

Previous studies have focused on comparing item-solutions versus parcel-solutions with respect to model fit and structural parameter estimates, and have focused on assessing the degree of within-allocation bias and *sampling variability* of parcel-solutions. In this study, we drew a distinction between two kinds of variability in parcel-solutions: *sampling variability*, referring to variability of solutions obtained across repeated samples given a single allocation of items to parcels, and *allocation variability,* referring to variability of solutions obtained across repeated allocations of items to parcels given a single sample. We showed that within-sample parcel-allocation variability has not been well studied—and not at all under the most highly recommended conditions for applying parceling (unidimensional, congeneric items for low *N* and low item communalities). In the few instances in which parcel-allocation variability has been investigated to date, circumscribed conditions were considered (Hall et al., 1999, Simulation 1; Sass & Smith, 2006, Study 1), leading to the premature conclusion that parcel-allocation variability in structural parameter estimates and model fit (for given numbers of items/factor, parcels/factor, items/parcel) is effectively nonexistent.

First, we extended MacCallum and Tucker's (1991) theoretical framework to show analytically that, even when items are unidimensional and congeneric in the population, alternative item-to-parcel allocations can affect model fit and structural parameter estimates in the *sample*—even though not in the population. Second, from this theoretical framework, we generated three hypotheses about the conditions under which parcel-allocation variability in model fit and structural parameter estimates should be most pronounced. Our simulation study employing repeated random allocations found support for each hypothesis. That is, parcel-allocation variability in model fit and structural parameter estimates is most pronounced for smaller samples, lower item communalities, and/or fewer number of parcels and items/parcel. Moreover, under these data conditions, the magnitude of parcel-allocation variability is practically concerning: a structural parameter estimate such as a factor correlation can range as much as .5 correlation units (but usually .1, .2, or .3) across allocations within such samples. And, for 14–44% of samples in these cells, > 5% allocations/sample switched statistical significance of the structural parameter estimate. Furthermore, model fit in a given sample varies substantially depending on the particular randomly chosen allocation (e.g., between 1 out of 4 and 1 out of 10 allocations within such samples differ in whether or not model fit is close). More generally, parcel-allocation variability in nonclose fit consistently approached zero only in cells where the proportion of samples with nonclose fit itself approached zero—a degenerate case. That is, anytime there was nonclose fit, there was meaningful parcel-allocation variability in that nonclose fit (see Figure 6). Our empirical example showed that these effects of parcel-allocation variability documented in the simulation—given no model error; a simple model; and unidimensional, congeneric, normal items—were conservative compared with what could be encountered in the real world when some of these idealized conditions do not hold.

Third, we related our findings on within-sample allocation variability to earlier findings on within-allocation sampling variability. In line with theoretical predictions, we replicated the results of Hau and Marsh (2004), Marsh et al. (1998), Nasser and Wisenbaker (2003), and Nasser-Abu and Wisenbaker (2006) in Panel 2 of Figures 4 and 5. Figure 5, Panel 2, showed that, within-allocation, across-sample averages of the structural parameter estimate had no bias. Figure 4, Panel 2, showed that, within-allocation, across-sample standard deviations of the structural parameter estimate decreased with increasing N and item communalities. Then, comparing parcel-allocation variability in the structural parameter estimate, *for each of 100 samples* with sampling variability in the structural parameter estimate, *for each of 100 allocations*, in a fully crossed design (Panel 1 vs. 2 in Figures 3 and 4), we showed that the same data conditions that create large sampling variability create large parcel-allocation variability. These conditions are small N, low item communalities, and fewer items/parcels. But, even under these conditions, Figure 5 showed that parcel-allocation variability in structural parameter estimates could be effectively ameliorated by statistically “removing” sampling error (Figure 5, Panel 2; in line with Hall et al., 1999, Simulation 1; Sass & Smith, 2006, Study 1). Regarding model fit, on the other hand, we showed that within a given cell of the simulation design, (a) sampling variability in model nonclose fit was quite similar for any allocation chosen, whereas (b) parcel-allocation variability in model nonclose fit varied considerably depending on the particular sample chosen. That is, model nonclose fit showed balanced clustering among samples within allocation (Figure 7) but unbalanced clustering among allocations within sample (Figure 6). But, again, if we statistically “removed” sampling error by averaging across samples within allocation, little/no allocation-to-allocation variability remained (Table 4, Panel 2; in line with Sass & Smith, 2006; Hall et al., 1999).

Implications of Different Patterns of Allocation Variability in Parameter Estimates Versus in Model Fit Statistics

Our simulation and empirical demonstration have important implications for the use of parceling in applied research. In practice, parceling is sometimes treated as an omnibus tool for ameliorating the effects of a variety of suboptimal data conditions (e.g., unreliable items, small samples, nonnormal items; Bagozzi & Edwards, 1998; Marsh et al., 1998) as well as for improving model fit compared with item solutions—as long as certain restrictions are met (e.g., unidimensional, congeneric items). We demonstrated that there are underappreciated costs associated with employing parceling under suboptimal data conditions: the introduction of an additional, nontrivial source of variability into model estimates. Investigators planning to employ parcels need be mindful of the sources and amount of sampling error in their analyses, which, in turn, fosters parcel-allocation variability in parameter and model fit estimates. In general, if parceling is to be used when sampling error is high (such as for $N \leq 150$ for a small model and/or item communalities $\leq .25$), we recommend that an applied researcher report and correct for parcel-allocation variability in model fit and parameter estimates. However, these cutoffs for reporting and correcting for parcel-allocation variability are not hard and fast for several reasons.

One reason these cutoffs are not hard and fast is that our simulation results evidenced nuanced compensatory trade-offs that are not captured in omnibus cutoffs. Moreover, these compensatory trade-offs differed for structural parameters versus model fit. An example trade-off is that as long as $N > 100$, medium-high communalities (.36) offset allocation variability for structural parameter estimates but not model fit. Another trade-off is that, as long as $N > 100$, more parcels (five rather than three) offset allocation variability for both structural parameter estimates and model fit. However, for few parcels (three), an N of 250 is needed—regardless of loading size—to offset allocation variability in model fit but not structural parameter estimates. In general, larger numbers of parcels was the most effective buffer of allocation variability in model fit whereas loading size was the most effective buffer of allocation variability in structural parameter estimates. Consequently, higher sample sizes and loading sizes can still be subject to meaningful allocation variability in model fit, even if they are not subject to meaningful allocation variability in structural parameter estimates. Researchers need be aware of the fact that parcel-allocation variability does not operate in the same way for structural parameter estimates as for model fit.

Another reason these cutoffs are not hard and fast is that they assume no or low measurement model error. Measurement model error (e.g., unmodeled cross loadings, unmodeled error covariances, misspecified number of factors) could create parcel-allocation variability at larger N's—(i.e., where sampling error is low). That is, in our model-error-free simulation we were able to eliminate parcel-allocation variability simply by reducing the amount of sampling error to a sufficient degree, but in the real world measurement model error could drive some parcel-allocation variability regardless of the amount of sampling error. However, characterizing the amount of parcel-allocation variability caused by commonly occurring forms of measurement model error was not a goal of this article.

Correcting for and Reporting Parcel-Allocation Variability: ParcelAlloc.sas Macro

Our results suggest a straightforward method for implementing our recommendation to report and control for parcel-allocation variability in model fit and parameter estimates when sampling error is high. Instead of reporting a single structural parameter estimate and model fit estimate from a single item-to-parcel allocation, a researcher would report the average structural parameter estimate, average analytic standard error, and average model fit estimate across an entire parcel-allocation distribution, and also would report how much these estimates are affected by parcel-allocation variability (i.e., the standard deviation of each estimate's parcel-allocation distribution). We also recommend reporting the percentage of allocations in which parameter estimates are found statistically significant. Examples of these reporting practices are: columns 2, 3, and 7 in Table 5.

Reporting the average structural parameter estimates and average model fit statistics serves to ensure that these estimates are not unduly affected by a chance, extreme allocation of items to parcels. In support of reporting the across-allocation mean of a structural parameter estimate, note that Figure 5, Panel 1, showed that the across-allocation mean is an unbiased estimate of its corresponding population parameter (.25), in the absence of model error. Reporting the standard deviation of the allocation distribution of these estimates clarifies for consumers of the parcel-analysis the amount of uncertainty introduced into the analysis through parceling. In support of reporting the standard deviation of an estimate's allocation distribution to communicate the magnitude of allocation variability, note that this logic parallels the accepted practice of reporting a standard error to communicate the magnitude of sampling variability.

To facilitate these recommended reporting practices, we have made available software tools (see http://www.unc.edu/∼ssterba/parcel.htm or contact S. Sterba) for creating many parcel-level data sets from a random item-to-parcel allocation distribution, and then analyzing and compiling results from the created set of parcel-level data sets. These software tools are compatible with any SEM model and also with missing data. A user submits an item-level data set with the variables to be included in the final statistical model. The user identifies how many items/parcel and parcels/factor are desired and how many random item-to-parcel allocations are desired. A SAS macro called *ParcelAlloc.sas* then performs random item-to-parcel allocations and generates a set of parcel-level data sets formatted to be read directly into the Monte Carlo utility of M*plus* (L. K. Muthén & Muthén, 1998–2008).[7] The user then specifies the SEM model as usual in M*plus* code, adding a few words of code that indicate to M*plus* to analyze not one parcel-level data set but the entire set of parcel-level data sets (e.g., adding the TYPE=MONTECARLO command). M*plus* then automatically provides model fit indices, parameter estimates, and analytic standard errors averaged across the entire parcel-allocation distribution, as well as the standard deviations of these distributions. M*plus* also automatically provides the percentage of allocations for which a parameter estimate was statistically significant. We have provided a detailed manual, along with three fully worked examples (including input files, output files, item-level data set, and parcel-level data sets) for different SEM and CFA models (see website or contact the first author).

Limitations and Future Directions

First, although the simulation and empirical example portions of this study focused on random rather than purposive item-to-parcel allocations, our analytic developments indicate that parcel-allocation variability will be a problem for purposive allocations as well.[8] To report and control for variability in purposive item-to-parcel allocations, researchers would first need to create a list of possible allocations using the chosen purposive method. This list would constitute the parcel-allocation distribution. A set of parcel-level data sets could be manually constructed from this list. These data sets could then be read into M*plus,* using the code provided, for analysis and automatic compilation across purposive allocations.

Second, our empirical example suggested that parcel-allocation variability increases under a variety of real-world conditions, such as model error and coarsely categorized data, which were not included in our simulation. The effects of such conditions on parcel-allocation variability need to be studied more systematically.

[7] The researcher does not have to analyze these parcel-level data sets in M*plus;* however, M*plus*'s Monte Carlo facility makes it easy to do so. A researcher could instead analyze these parcel-level data sets in batch mode, in other software programs, and compile the results independently.

[8] Of historical interest, we note a relation between the present research on parcel-allocation variability and early concern regarding variation in split half reliability across alternate splits of test items. Specifically, the former is analogous to taking a set of items, forming two parcels, and looking at the variation in the correlation between those two parcels. Cronbach (1951, pp. 309–311) discussed topics similar to those covered here, such as random versus purposive splitting, obtaining the mean of the distribution of all possible splits (i.e., coefficient alpha), and the fact that smaller variability was found within-sample across splits than within-split across-sample. Further connections are possible but are not pursued here.

Third, our simulations used 100 allocations per sample to conserve computing time. Selected checks with 500 allocations per sample yielded the same overall pattern of simulation results. Future research is needed to investigate how large the parcel-allocation distribution needs to be in order to obtain a stable allocation distribution standard deviation and range for a single sample—given size of the model, number of items, and number of parcels. In practice, researchers with a single sample should ensure that if they incremented their chosen number of allocations (e.g., from 100 to 150) their results remain stable; if they do not, the number of allocations should be increased.

Fourth, space did not permit a detailed presentation of parcel allocation variability in standard errors. Standard errors of parameter estimates vary across allocations within sample, just as parameter estimates do. Such parcel allocation variability in standard errors can be inferred from the varying statistical significance rates across allocations in Tables 2 and 5. For further details, we have provided an online appendix with plots similar to Figures 3, 4, and 5 but for analytic standard errors of the factor correlation rather than the correlation estimate itself (see website or contact the first author).

CONCLUSIONS

Many researchers currently use parceling (Bandalos & Finney, 2001)—presumably often under low item communalities and low N, as is commonly recommended. We showed that under these conditions, the amount of within-sample parcel-allocation variability in structural parameter estimates and model fit can be concerning—enough to alter substantive conclusions. Under these circumstances, we recommend using our software tools to report the magnitude of parcel-allocation variability and to minimize the effects of allocation-to-allocation variability on substantive conclusions.

ACKNOWLEDGMENTS

This work was supported by an NIMH F31 MH080494 award to S. Sterba.

REFERENCES

Asendorpf, J. B., & Wilpers, S. (1998). Personality effects on social relationships. *Journal of Personality and Social Psychology, 74,* 1531–1544.

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 (Eds.). *New developments and techniques in structural equation modeling* (pp. 269–297). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Benet-Martinez, V., & Karakitapoglu, Z. (2003). The interplay of cultural syndromes and personality in predicting life satisfaction: Comparing Asian Americans and European Americans. *Journal of Cross-Cultural Psychology, 34,* 38–60.

Cattell, R. B. (1956). Validation and intensification of the sixteen personality factor questionnaire. *Journal of Clinical Psychology, 12,* 205–214.

Cattell, R. B. (1974). Radial parcel factoring vs. item factoring in defining personality structure in questionnaires: Theory and experimental checks. *Australian Journal of Psychology, 26,* 103–119.

Clara, I. P., Cox, B. J., & Enns, M. W. (2003). Hierarchical models of personality and psychopathology: The case of self-criticism, neuroticism and depression. *Personality and Individual Differences, 35,* 91–99.

Coffman, D. L., & MacCallum, R. C. (2005). Using parcels to convert path analysis models into latent variable models. *Multivariate Behavioral Research, 40*, 235–259.

Cohen, S., & Hoberman, H. M. (1983). Positive events and social supports as buffers of life change stress. *Journal of Applied School Psychology, 13,* 99–125.

Costa, P. T., & McCrae, R. R. (1985). *The NEO Personality Inventory: Manual (Form S and Form R).* Odessa, FL: Psychological Assessment Resources, Inc.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

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

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.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

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. (2000). *Assessment of item parcels in representing latent variables* (Unpublished doctoral dissertation). Athens, GA: University of Georgia.

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.

Kishton, J. M., & Widaman, K. F. (1994). Unidimensional versus domain representative parceling of questionnaire items: An empirical example. *Educational and Psychological Measurement, 54,* 757–765.

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.

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, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9,* 151–173.

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.

MacCallum, R. C., Widaman, K. F., Preacher, K. J., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Psychological Methods, 4,* 611–637.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4,* 84–89.

Marsh, H. W., Hau, K.-T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33,* 181–220.

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

Muthén, B. O., du Toit, S. H., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Conditionally accepted for publication in Psychometrika.

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

Nasser, F., & Wisenbaker (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.

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

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.

Saucier, G. (2002). Orthogonal markers for orthogonal factors: The case of the Big Five. *Journal of Research in Personality, 36,* 1–31.

Sterba, S. K., Egger, H. L., & Angold, A. (2007). Diagnostic specificity and non-specificity in the dimensions of preschool psychopathology. *Journal of Child Psychology and Psychiatry, 48,* 1005–1013.

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.

Yuan, K.-H., Bentler, P., & Kano, Y. (1997). On averaging variables in a confirmatory factor model. *Behaviormetrika, 24,* 71–83.