Psychological Methods
2017, Vol. 22, No. 1, 47–68


1082-989X/17/$12.00 http://dx.doi.org/10.1037/met0000067

# Effects of Parceling on Model Selection: Parcel-Allocation Variability in Model Ranking

Sonya K. Sterba and Jason D. Rights
Vanderbilt University

Research interest often lies in comparing structural model specifications implying different relationships among latent factors. In this context parceling is commonly accepted, assuming the item-level measurement structure is well known and, conservatively, assuming items are unidimensional in the population. Under these assumptions, researchers compare competing structural models, each specified using the same parcel-level measurement model. However, little is known about consequences of parceling for model selection in this context—including whether and when model ranking could vary across alternative item-to-parcel allocations within-sample. This article first provides a theoretical framework that predicts the occurrence of parcel-allocation variability (PAV) in model selection index values and its consequences for PAV in ranking of competing structural models. These predictions are then investigated via simulation. We show that conditions known to manifest PAV in absolute fit of a single model may or may not manifest PAV in model ranking. Thus, one cannot assume that low PAV in absolute fit implies a lack of PAV in ranking, and vice versa. PAV in ranking is shown to occur under a variety of conditions, including large samples. To provide an empirically supported strategy for selecting a model when PAV in ranking exists, we draw on relationships between structural model rankings in parcel- versus item-solutions. This strategy employs the across-allocation modal ranking. We developed software tools for implementing this strategy in practice, and illustrate them with an example. Even if a researcher has substantive reason to prefer one particular allocation, investigating PAV in ranking within-sample still provides an informative sensitivity analysis.

*Keywords:* parceling, model selection, structural equation modeling, parcel-allocation variability, model ranking

*Supplemental materials:* http://dx.doi.org/10.1037/met0000067.supp

Social scientists frequently *parcel* (average or sum subsets of items) and use the parcel scores as factor indicators. Recent reviews indicate that parceling is used in 20%, 17%, 63%, and 44% of structural equation modeling applications, across various journals (Hall, Snell, & Foust, 1999; Bandalos & Finney, 2001; Plummer, 2000; Williams & O'Boyle, 2008). Whereas item-level analyses are usually preferable, in the context of complex models but modest sample sizes, item-level analyses may become impractical or encounter estimation problems. Parceling is one of few alternatives available in this context (see Bagozzi & Edwards, 1998; Coffman & MacCallum, 2005; Hau & Marsh, 2004; Little, Rhemtulla, Gibson, & Schoemann, 2013; Marsh, Lüdtke, Nagengast, Morin, & von Davier, 2013; Matsunaga, 2008; Meade & Kroustalis, 2006; Nasser & Wisenbaker, 2003; Sass & Smith, 2006; Yang, Nay, & Hoyle, 2010).

But parceling is not universally applicable. Because parceling can obscure the detection of measurement model misspecifications such as unmodeled multidimensionality, parceling has often been considered most defensible when items per parcel are unidimensional in the population (i.e., items per parcel load on one and the same factor; see Bandalos, 2002, 2008; Hall et al., 1999; Hau & Marsh, 2004; Landis, Beale & Tesluk, 2000; Little, Cunningham, Shahar & Widaman, 2002; Marsh & O'Neill, 1984; Marsh et al., 2013; Matsunaga, 2008; Meade & Kroustalis, 2006; Plummer, 2000; Rogers & Schmitt, 2004; Sass & Smith, 2006; Yang et al., 2010; Yuan, Bentler, & Kano, 1997). Parceling has been discouraged when research interest lies in scale development or exploring the number of factors; rather, parceling is typically implemented when interest is in structural relations among factors, under the assumption of a known population item-level measurement model (e.g., Bandalos, 2002; Bandalos & Finney, 2001; Little et al., 2002, 2013; Marsh et al., 2013; Matsunaga, 2008; Meade & Kroustalis, 2006; Nasser-Abu & Wisenbaker, 2006; Plummer, 2000; Rogers & Schmitt, 2004; Sass & Smith, 2006; Stucky, Gottfredson, & Panter, 2012; Williams & O'Boyle, 2008). To increase the plausibility of this assumption, parceling has been considered acceptable only when latent constructs are well defined theoretically (e.g., Little et al., 2013; Matsunaga, 2008) and preferably have been subjected to previous item-level analyses investigating dimensionality (e.g., Bandalos & Finney, 2001; Marsh et al., 2013).

This article was published Online First January 25, 2016.

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

We thank Robert C. MacCallum and Kristopher J. Preacher for helpful comments on previous versions of this article.

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

For a given sample, there are typically thousands of possible ways a researcher could allocate items to parcels—given a prespecified number of parcels per factor and items per parcel. These are sometimes called alternative parceling strategies but are here called alternative *parcel allocations*. If we were to repeatedly, randomly allocate items to parcels for a given number of parcels/factor and items/parcel, we would eventually, by chance, employ many purposive parceling strategies in existence (e.g., balancing, correlational, factorial, odd–even, adjacent-loading, etc.). Thus, we can consider these strategies as special case instantiations of random item-to-parcel allocations. When evaluating the fit of a single model in isolation, it is known that absolute model fit can vary meaningfully (e.g., from poor to excellent) across repeated allocations of items to parcels—even when items are unidimensional in the population (Sterba, 2011; Sterba & MacCallum, 2010). Such *parcel-allocation variability* (PAV) in absolute fit within-sample can arise simply due to sampling error (but would be exacerbated in the context of model error in the measurement model). In this light, it may be unsettling to select and report absolute fit based on just one item-to-parcel allocation. On one hand, a particular item-to-parcel allocation could be substantively justified by the researcher. On the other hand, it is possible that the researcher's substantive justification is underdeveloped or even wrong. Hence, it is more reassuring to know a range of possible absolute fit results from alternative item-to-parcel allocations within that researcher's single sample. Under some data conditions, this range will be narrow, such that substantive conclusions about model adequacy would not be contingent on the allocation chosen. Under other conditions, this range can be broad. Building on MacCallum and Tucker (1991) and Bandalos (2002); Sterba and MacCallum (2010) provided a theoretical framework that identified conditions increasing PAV in absolute fit of a single parcel-level model.

However, frequently researchers are not simply interested in a single parcel-level model. Rather, researchers often parcel with the goal of comparing competing structural specifications that imply different relationships among latent factors, assuming a known item-level measurement structure. In other words, researchers who parcel are often interested in *model selection* among different *structural model* specifications, where the parcel-level measurement model is the same across the set of candidate models. For instance, this was the focus of many recent applications, including Booth, Murray, Marples, and Batey (2013); Daspit, Tillman, Boyd, and McKee (2013); Dunkley, Ma, Lee, Preacher, and Zuroff (2014); Flack, Salmivalli, and Idsoe (2011); Gallagher, Lopez, and Preacher (2009); Geiser, Keller, and Lockhart (2013); Gellert, Ziegelmann, and Schwarzer (2012); Hankonen, Konttinen, and Absetz (2014); Jackson and Gaertner (2010); Kuhn and Holling (2009); Liao, O'Brien, Jimmieson, and Restubog (2015); Mairet, Boag, and Warburton (2014); Malmberg and Little (2007); Martin et al. (2011); Nouwen, Urquhart Law, Hussain, McGovern, and Napier (2009); Owuamalam, Issmer, Zagefka, Klaben, and Wagner (2014); Segrin, Woszidlo, Givertz, and Montgomery (2013); Sierau and Herzberg (2012); Winkler, Busch, Clasen, and Vowinkel (2015); Zampetakis, Kafetsios, Bouranta, Dewett, and Moustakis (2009); and Zheng, Gaumer Erickson, Kingston, and Noonan (2014). Here, only the structural specification differs across models. Each competing structural model corresponds with a different substantive theory. Among the structural models under comparison, none may represent the population-generating process exactly. Model selection indices may be used to, for instance, select the candidate model that is closest to the generating structural process or select the most generalizable model (Burnham & Anderson, 2004; Myung & Pitt, 1998).

Despite the frequent application of parceling in the context of structural model selection, we know little about its consequences—in terms of whether and when model selection results (e.g., model ranking) could vary across alternative item-to-parcel allocations within-sample. This article is devoted to filling this gap. This article makes the following contributions. First, we extend the theoretical framework of Sterba and MacCallum (2010)—which allowed for sampling error in a parcel-level measurement model—to include a structural model, as well as sampling and model error in the structural model. Then we apply this framework to the context of model selection. We use this extended framework to identify situations under which PAV in model selection index values can occur when comparing structural models. When PAV in selection index values can occur, the framework is used to predict conditions that increase risk of PAV in model ranking. Second, we investigate these predictions in a simulation. We show that the conditions under which there is PAV in model ranking within-sample may be similar to *or* greatly different from conditions under which previous research found PAV in absolute model fit. In other words, it is not sufficient to assume that low PAV in absolute model fit implies a lack of PAV in model ranking, and vice versa. Third, when PAV in model ranking exists, we provide an empirically supported strategy for deciding what model to select, which draws on the relationship (shown later) between model ranking in parcel- versus item-solutions. Finally, we provide software tools for implementing this strategy in applied practice and demonstrate their implementation in an empirical example.

Before proceeding, however, we must clarify that although the topic of whether to parcel has recently been framed in terms of *pro* and *con* arguments (e.g., Little et al., 2013; Marsh et al., 2013; Matsunaga, 2008), previous authors found some common ground on the admissibility of parceling in particular situations—including the exact context considered in this article (i.e., interest in structural relations, unidimensional items in the population, modest *N*). The present study contains empirical results that stand on their own and are in essence agnostic to a particular side. That is, this study presents knowledge about the consequences of parceling within-sample in an understudied but common setting—model selection. As conceptualized here, the *existence* of PAV in model selection results (or any other results) itself needn't discredit nor encourage the use of parceling. At issue here is this: *When* parceling is used, how can we understand and investigate PAV in model selection within-sample and interpret results in this light?

## Theoretical Framework

In this section, we begin by representing a sample item-level covariance structural model. Next, we relate this expression to a sample parcel-level covariance structure model. Consistent with the common context for parceling—moderate sample size and an assumed-known item-level measurement model—our model representation allows only for sources of sampling error and structural-model-specific model error. That is, we assume no model error arises from the measurement model—an assumption

made implicitly or explicitly by most researchers who employ and/or recommend parceling (see Little et al., 2002, for review).[1] Finally, we then detail implications of this framework for PAV in model selection.

## Item-Level Covariance Structure Model in the Sample

We start by considering a covariance structure model for an item-level *sample* covariance matrix, $\mathbf{S}_i$. An *i* subscript denotes *item level*. We can consider the illustrative situation where items per factor are unidimensional with unit variances in the population so that our later discussion can refer interchangeably to the impact of loading or communality size (the latter is the square of the former).

There are *m* items measuring *q* factors. In Equation (1), $\mathbf{\Lambda}_i$ is a $m \times q$ common factor loading matrix, $\mathbf{C}_{cc_i}$ is a $q \times q$ common factor covariance matrix, and $\mathbf{\Psi}_i$ is a $m \times m$ diagonal matrix of unique factor loadings.

$$\mathbf{S}_i = \mathbf{\Lambda}_i \mathbf{C}_{cc_i} \mathbf{\Lambda}_i' + \mathbf{\Psi}_i^2 + \mathbf{\Delta}_{SE1_i} \tag{1}$$

Following MacCallum and Tucker (1991) and MacCallum, Widaman, Zhang, and Hong, (1999), the term $\mathbf{\Delta}_{SE1_i}$ represents all lack of fit due to sampling error arising from the measurement model. The sources of sampling error contributing to $\mathbf{\Delta}_{SE1_i}$ are represented in greater detail in the Appendix A equations (see also MacCallum, 2013; MacCallum et al., 1999; Bandalos, 2002; Bandalos & Finney, 2001; Matsunaga, 2008; Meade & Kroustalis, 2006).

In Equation (2), we introduce a structural model which allows constraints on common factor covariances in $\mathbf{C}_{cc_i}$. In Equation (2), $\mathbf{\Omega}_i$ is a $q \times q$ matrix of residual covariances among common factors and $\mathbf{B}_i$ is a $q \times q$ matrix of regressions among common factors.

$$\mathbf{C}_{cc_i} = (\mathbf{I} - \mathbf{B}_i)^{-1} \mathbf{\Omega}_i (\mathbf{I} - \mathbf{B}_i')^{-1} + \mathbf{\Delta}_{SE2_i} + \mathbf{\Delta}_{ME_i} \tag{2}$$

In MacCallum and Tucker's (1991) original framework, $\mathbf{C}_{cc_i}$ was unstructured; no lack of fit due to sampling error could arise from such a saturated structural model because it has no constraints. In contrast, the constraints imposed on $\mathbf{C}_{cc_i}$ in Equation (2) can cause misfit; indeed, sampling error alone could cause the constraints to be inappropriate even if they hold in the population. Lack of fit due to sampling error, arising from the structural model, is represented by $\mathbf{\Delta}_{SE2_i}$ in Equation (2). The Appendix A equations represent in greater detail the sources of sampling error contributing to $\mathbf{\Delta}_{SE2_i}$.

Finally, the term $\mathbf{\Delta}_{ME_i}$ in Equation (2) represents all lack of fit from model error specifically in the structural model. In MacCallum and Tucker (1991); MacCallum, Widaman, Preacher, and Hong (2001), and MacCallum (2013), model error was not specific to the structural model, as it is here. This model error might arise due to, for example, parametric misspecifications of the population structural model.

In Equation (3), we substitute Equation (2) into (1) and denote $\mathbf{\Delta}_{SE_i} = \mathbf{\Lambda}_i \mathbf{\Delta}_{SE2_i} \mathbf{\Lambda}_i' + \mathbf{\Delta}_{SE1_i}$ to yield a reduced-form item-level covariance structure model in the sample:

$$\mathbf{S}_i = \mathbf{\Lambda}_i (\mathbf{I} - \mathbf{B}_i)^{-1} \mathbf{\Omega}_i (\mathbf{I} - \mathbf{B}_i')^{-1} \mathbf{\Lambda}_i' + \mathbf{\Psi}_i^2 + \mathbf{\Lambda}_i \mathbf{\Delta}_{ME_i} \mathbf{\Lambda}_i' + \mathbf{\Delta}_{SE_i}. \tag{3}$$

## Parcel-Level Covariance Structure Model in the Sample

Having first considered the item-level covariance structure model in the sample, we now turn our attention to the parcel-level covariance structure model in the sample. Let $\mathbf{A}$ be an $m \times j$ selection matrix. It serves to allocate *m* items to *j* parcels in the measurement model for a given, prespecified number of parcels/factor and items/parcel, within a single sample. Allocation $\mathbf{A}$ could be chosen randomly or purposively from the within-sample distribution of potential allocations that has the desired number of parcels/factor and items/parcel. A *p* subscript indicates parcel level. The parcel-level covariance structure model can be obtained by pre- and postmultiplying Equation (3) by $\mathbf{A}$ and $\mathbf{A}'$, respectively:

$$\mathbf{S}_p = \mathbf{A}\mathbf{\Lambda}_i((\mathbf{I} - \mathbf{B}_i)^{-1} \mathbf{\Omega}_i (\mathbf{I} - \mathbf{B}_i')^{-1}) \mathbf{\Lambda}_i' \mathbf{A}' + \mathbf{A}\mathbf{\Lambda}_i \mathbf{\Delta}_{ME_i} \mathbf{\Lambda}_i' \mathbf{A}' + \mathbf{A}\mathbf{\Psi}_i^2 \mathbf{A}' + \mathbf{A}\mathbf{\Delta}_{SE_i} \mathbf{A}'. \tag{4}$$

Following Sterba and MacCallum (2010), MacCallum (2013), and Sterba (2011), for the parcel-level model, we can define the following specific to allocation $\mathbf{A}$: loadings, $\mathbf{\Lambda}_p = \mathbf{A}\mathbf{\Lambda}_i$, unique variances, $\mathbf{\Psi}_p^2 = \mathbf{A}\mathbf{\Psi}_i^2\mathbf{A}'$, and lack of fit due to sampling error, $\mathbf{\Delta}_{SE_p} = \mathbf{A}\mathbf{\Delta}_{SE_i}\mathbf{A}'$. Additionally, we newly define $\mathbf{\Delta}_{ME_p} = \mathbf{\Lambda}_p \mathbf{\Delta}_{ME_i} \mathbf{\Lambda}_p'$ as the lack of fit due to structural model error that is specific to allocation $\mathbf{A}$. The latter expression shows that the contribution of structural model error is specific to allocation $\mathbf{A}$ because its contribution is differentially weighted by the allocation-specific loadings. Simplifying Equation (4) using these definitions yields Equation (5).

$$\mathbf{S}_p = \mathbf{\Lambda}_p((\mathbf{I} - \mathbf{B}_i)^{-1} \mathbf{\Omega}_i (\mathbf{I} - \mathbf{B}_i')^{-1}) \mathbf{\Lambda}_p' + \mathbf{\Psi}_p^2 + \mathbf{\Delta}_{ME_p} + \mathbf{\Delta}_{SE_p} \tag{5}$$

Although the same structural parameters $\mathbf{B}_i$ and $\mathbf{\Omega}_i$ appear in both the item-level expression (Equation [3]) and parcel-level expression (Equation [5]), their sample estimates will *not* be the same across allocations within-sample because $\mathbf{\Delta}_{SE_p}$ and $\mathbf{\Delta}_{ME_p}$ have allocation-specific impacts on structural estimates.

We now denote the population model-implied parcel-level covariance structure as $\tilde{\mathbf{\Sigma}}_p$, where $\tilde{\mathbf{\Sigma}}_p = \mathbf{\Lambda}_p(\mathbf{I} - \mathbf{B}_i)^{-1}\mathbf{\Omega}_i(\mathbf{I} - \mathbf{B}_i')^{-1}\mathbf{\Lambda}_p' + \mathbf{\Psi}_p^2$. In other words, $\tilde{\mathbf{\Sigma}}_p$ is a function of all model parameters and here is evaluated at population values of these parameters (i.e., values of the parameters obtained if the model were fit in the population). Similar to MacCallum et al. (2001, Equation [15]), we substitute $\tilde{\mathbf{\Sigma}}_p$ into Equation (5) to yield Equation (6).

$$\mathbf{S}_p = \tilde{\mathbf{\Sigma}}_p + \mathbf{\Delta}_{ME_p} + \mathbf{\Delta}_{SE_p} \tag{6}$$

In Equation (6), the parcel-level sample covariance matrix is partly accounted for by the population model-implied covariance structure and partly explained by misfit due to structural model error and sampling error.

[1] Violation of this assumption would increase PAV in a manner similar to reducing sample size; see Discussion section.

## Model Selection Between Competing Structural Specifications of Parcel-Level Models

Suppose now that we have not one but two such parcel-level models, designated Models *a* and *b*. Models *a* and *b* differ from each other only in the structural submodel. That is, the measurement submodel specification for Models *a* and *b* are the same. Models *a* and *b* are fit in the sample using the same parcel-allocation.

Further suppose that we want to select between parcel-level Models *a* and *b*. For this purpose, we might consider computing a likelihood ratio test (LRT) statistic or computing differences in information criteria.[2] First, consider the LRT statistic, here referred to as *T*. Because the LRT will require nested models, here suppose Model *a* is nested within *b*. When all items are assumed mean deviated, *T* can be expressed as Equation (7), as explained in Appendix B.

$$T = (N-1)\Big(\ln\big|\hat{\Sigma}_p^a\big| - \ln\big|\hat{\Sigma}_p^b\big| + tr\Big(\hat{\Sigma}_p^{a-1}\Big(\tilde{\Sigma}_p^a + \Delta_{ME_p}^a + \Delta_{SE_p}^a\Big)\Big) - tr\Big(\hat{\Sigma}_p^{b-1}\Big(\tilde{\Sigma}_p^b + \Delta_{ME_p}^b + \Delta_{SE_p}^b\Big)\Big)\Big) \quad (7)$$

In Equation (7), superscripts *a* or *b* refer to Models *a* or *b*, respectively. Also, $\hat{\Sigma}_p$ refers to a sample model-implied parcel-level covariance structure. It is important to note that Equation (7) is not a computational formula for *T* that would be used in empirical practice (see later Simulation Method section for a computational formula). Rather, Equation (7) is useful for understanding how *T* is affected by sampling and model error, as described in the next section.

Researchers also often use differences in Bayesian information criteria (Schwarz, 1978) or Akaike's information criterion (Akaike, 1973) to aid in model selection: $\Delta\text{BIC} = T + \ln N(k^a - k^b)$ and $\Delta\text{AIC} = T + 2(k^a - k^b)$. Here, $k^a - k^b$ is the difference in the number of free parameters between Models *a* and *b*. In the next section we use our framework to infer whether within-sample PAV in model *selection index values* (i.e., *T* or ΔBIC or ΔAIC) is possible under three situations. Note that we are *not* yet discussing model ranking. For a given pair of models, the standard deviation of the within-sample PAV distribution of *T,* ΔBIC, and ΔAIC will be the same because these indices each differ by a constant. Hence, the next section can refer to all three index values generically.

## Parcel-Allocation Variability in Model Selection Index Values Within-Sample

Here we consider three key situations and we use the above framework to determine whether PAV in model selection index values can occur in each situation. We define *structural model differences* as between-model differences in inappropriate constraints in the structural model and/or between-model differences in superfluous parameters in the structural model. Structural model differences are represented as $(\Delta_{ME_p}^b - \Delta_{ME_p}^a)$.

**Case I: Models *a* and *b* are equivalent models and there is sampling error.** First, suppose structural model differences between Models *a* and *b* approach 0 [i.e., $(\Delta_{ME_p}^b - \Delta_{ME_p}^a) \rightarrow \mathbf{0}$]. This implies that $(\tilde{\Sigma}_p^b - \tilde{\Sigma}_p^a) \rightarrow \mathbf{0}$ because Models *a* and *b* both were defined as having no model error in the corresponding item-level measurement models. It also follows that $(\Delta_{SE_p}^b - \Delta_{SE_p}^a) \rightarrow \mathbf{0}$ and $(\hat{\Sigma}_p^b - \hat{\Sigma}_p^a) \rightarrow \mathbf{0}$. This is because, although sampling error could manifest differently across models, as the models become more and more similar, the sampling error must eventually have the same manifestation per model. Now, consider the limiting situation where Models *a* and *b* are *equivalent models* in the sense of Lee and Hershberger (1990) and MacCallum, Wegener, Uchino, and Fabrigar (1993); this corresponds with structural model differences between Models *a* and *b* being 0 in the sample. In this limiting case, sampling error alone *cannot induce PAV in model selection* index values in Equation (7). Indeed, in this case Equation (7) would be 0.

**Case II: Models *a* and *b* have structural differences and there is no sampling error.** Suppose sampling error approaches 0 (i.e., $\Delta_{SE_p}^b \rightarrow \mathbf{0}$ and $\Delta_{SE_p}^a \rightarrow \mathbf{0}$), meaning the final term in Equation (5) drops out for each model and that $\hat{\Sigma}_p^a \rightarrow \tilde{\Sigma}_p^a$ and $\hat{\Sigma}_p^b \rightarrow \tilde{\Sigma}_p^b$ (see Appendix B). If item-level loadings within-factor are *un*equal in the population, then the impact of structural model error (present in Model *a,* $\Delta_{ME_p}^a = \Lambda_p^a \Delta_{ME_i}^a \Lambda_p^{a\prime}$ and/or Model *b,* $\Delta_{ME_p}^b = \Lambda_p^b \Delta_{ME_i}^b \Lambda_p^{b\prime}$) is still weighted by allocation-specific matrices (i.e., $\Lambda_p' = \mathbf{A}\Lambda_i$). This implies the possibility of PAV in model selection index values.[3] However, if item-level loadings within-factor are equal in the population, then the impact of structural model error (present in Model *a* or *b* or both) is no longer allocation-specific because its contribution is no longer weighted by allocation-specific matrices. This implies no PAV in model selection index values.[4]

**Case III. Models *a* and *b* have structural differences and there is sampling error.** Under Case III, neither $\Delta_{SE_p}^a$ nor $\Delta_{SE_p}^b$ drop out of Equation (7). Additionally, the influence of structural model error (present in Model *a,* $\Delta_{ME_p}^a$, and/or Model *b,* $\Delta_{ME_p}^b$) is allocation-specific because, even if item-level loadings within-factor are equal in the population, they will be unequal in the sample due to sampling error. Thus, the combination of sampling error plus structural model differences between *a* and *b* can yield PAV in values of a model selection index such as the LRT statistic, *T*.

[2] We refer to LRT and information criteria all as "selection indices" here because we use them later for the same purpose—to rank models (Maxwell & Delaney, 2004; Rodgers, 2010; Sterba & Pek, 2012)—despite the former coming from a distinct hypothesis-testing tradition.

[3] This point predicted by the theoretical framework was verified by a small simulation. A population item-level dataset was generated with unidimensional items per factor and unequal item-level loadings per factor. The item-level dataset was generated in SAS Proc IML using Kaiser and Dickman's Method 2 (Kaiser & Dickman, 1962, Equation 5, p. 180) to have no sampling error (meaning that fitting the generating item-level model in the sample would exactly reproduce the population correlation matrix). The nine items per factor were repeatedly randomly allocated to three three-item parcels, and two parcel-level models were fit with the same measurement specification but different structural specifications. PAV in model selection index values was observed.

[4] This point predicted by the theoretical framework was verified by a small simulation. The same procedures from Footnote 3 were followed to generate a population item-level dataset (i.e., no sampling error) with unidimensional items per factor and *equal* item-level loadings per factor. The nine items per factor were again repeatedly randomly allocated to three three-item parcels, and two parcel-level models were fit with the same measurement specification but different structural specifications. PAV in model selection index values was *not* observed.

Substantively, however, our interest is not simply in PAV in *T,* ΔBIC, ΔAIC, or values of another selection index. Rather, we are interested in PAV in the decision regarding which model to retain as best fitting—that is, PAV in *model ranking*. This is the topic of the next section.

## Parcel-Allocation Variability in Model Ranking Within-Sample

*PAV in model ranking* is defined as when the ranking preference flips from model *a* to *b* or *b* to *a* in a nonzero proportion of allocations, within-sample. Sampling error plus structural model differences can induce PAV in fit ranking within-sample, but are not sufficient to guarantee that it will arise. In a given parcel-allocation within sample, model ranking of *b* over *a* (i.e., support for Model *b*) corresponds to $T(df) > T_{crit}(df)$ for the LRT statistic, $\Delta BIC > 0$ for the Bayesian information criterion, or $\Delta AIC > 0$ for Akaike information criterion. Here, $df = k^b - k^a$. $T_{crit}(df)$ is the critical value of the LRT statistic at that $df$. The null hypothesis for the LRT is that there is no difference in the fit of Models *a* and *b* in the population. Unlike the LRT, ΔBIC, and ΔAIC do not require that Model *a* be nested in Model *b*.

Next, we describe two hypotheses involving conditions expected to increase or decrease the risk of PAV in model fit ranking within-sample, in this context. Figure 1 illustrates these two hypotheses. In the subsequent section, a simulation is used to evaluate each hypothesis.

## Hypotheses

*Hypothesis 1:* PAV in model ranking can arise at small, medium, or large structural differences, depending on the amount of sampling error; it is more likely to arise when structural differences yield a selection index value closer to its decision threshold.

Consider that each selection index has a *decision threshold* that distinguishes support for Model *a* versus *b*. In the case of the LRT, this threshold is $T_{crit}(df)$, and for ΔBIC or ΔAIC, this threshold is 0. The across-allocations within-sample distribution of the selection index's values (i.e., *T,* ΔBIC, or ΔAIC) needs to overlap that index's decision threshold in order for PAV in model ranking to arise. As an example, if ΔBIC ranges from 10 to 15 across allocations within-sample, this implies no PAV in model ranking, because the decision threshold for ΔBIC is 0 and all allocations prefer Model *b*. On the other hand, if ΔBIC ranges from −3 to 4 across allocations within-sample, overlapping the decision threshold, there would be PAV in model ranking in that sample because some allocations prefer Model *b* and some prefer Model *a*. Alternatively, if ΔBIC ranges from −9 to −2 across allocations within-sample, there would be no PAV in model ranking in that sample because all allocations prefer Model *a*.

In the Figure 1 heuristic illustration, each distribution represents an across-allocation *within-sample* distribution of a generic model selection index (ΔBIC, ΔAIC, or *T*). The vertical line represents

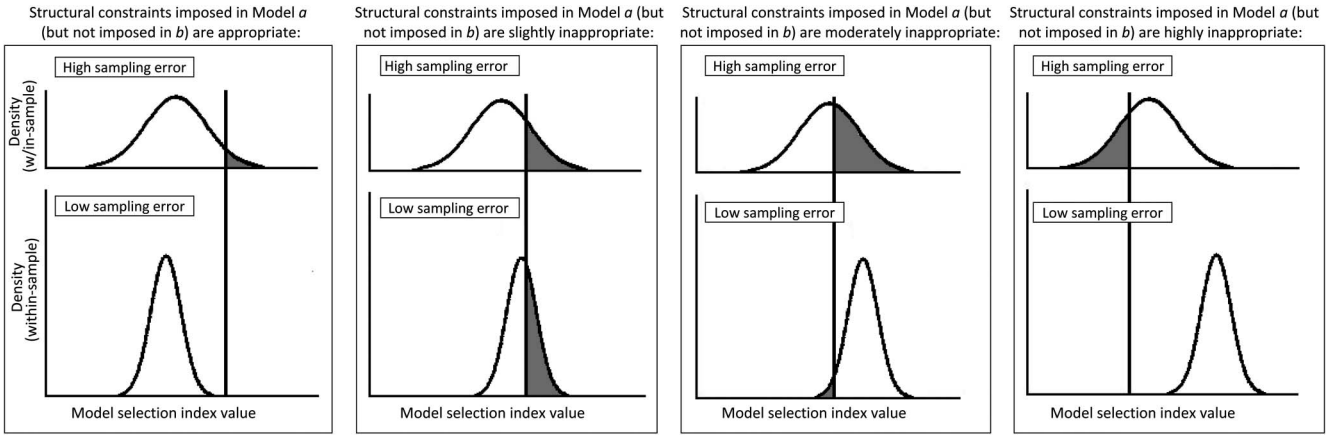


*Figure 1.* Illustration of Hypotheses 1 and 2. Hypothesis 1 is illustrated by comparing across the four panels (columns). Hypothesis 2 is illustrated by comparing top and bottom distributions within each panel. Each distribution is a within-sample across-allocation distribution of model selection index values obtained when comparing Models *a* and *b*. The model selection index could be either change in Bayesian information criterion (ΔBIC), change in Akaike information criterion (ΔAIC), or likelihood ratio test statistic; see text for definitions. The vertical line represents the decision threshold for a model selection index. To the right of the decision threshold, Model *b* (the more complex model) is preferred; to the left of the decision threshold, Model *a* is preferred. In this illustration, the decision threshold stays at the same location in all plots for a given index because Model *a* and *b* have the same $\Delta df$ in all plots (as will be the case later in the simulation). *Incidence* of parcel-allocation variability in model ranking occurs when the decision threshold overlaps the across-allocation distribution, within-sample. *Magnitude* of parcel-allocation variability in model ranking within sample is denoted by the size of the shaded portion of the distribution (i.e. the proportion of allocations preferring the least favored model, within-sample). Each column corresponds with a different amount of structural differences between Model *a* and *b.* High manifestation of sampling error corresponds with lower sample size and lower communalities.

the decision threshold for that index; it stays in the same location for all panels of Figure 1 (see Figure 1 notes). To the right of the decision threshold, Model *b* (the more complex model) is preferred and to the left of the decision threshold, Model *a* is preferred. When structural differences ($\boldsymbol{\Delta}^{b}_{ME_p} - \boldsymbol{\Delta}^{a}_{ME_p}$) lead on average to selection index values closer to its decision threshold, there is a greater potential for that across-allocations within-sample distribution to overlap the decision threshold. This overlap corresponds with the *incidence* of PAV in model ranking. Note that, when PAV in model ranking occurs, the size of the shaded area depicts *magnitude* of PAV– the proportion of allocations preferring the least favored model, within-sample.

Hypothesis 1 additionally states that the risk of PAV in model ranking occurring should not be limited to a particular degree of structural differences between models. This is because, for a given degree of structural difference between models in the population, the across-allocation within-sample distribution of the selection index could be closer to its decision threshold (higher risk of PAV in ranking) or farther from its decision threshold (lower risk of PAV in ranking) depending on the amount of sampling error. In Figure 1, the four panels (columns) correspond with four different degrees of structural differences between Models *a* and *b* in the population. Figure 1 illustrates that PAV in ranking should occur under some conditions within each panel (further discussed in Hypothesis 2, below). Specifically, Figure 1 illustrates the expectation that, under certain conditions, PAV in ranking can be encountered when structural constraints imposed on Model *a* (but not imposed on Model *b*) are highly inappropriate, moderately appropriate, slightly inappropriate, or appropriate. Only when all allocations within sample prefer a single model (e.g., bottom right plot of Figure 1), should there be no PAV in ranking.

Recall that the Figure 1 illustration is shown for a single model selection index. Because each selection index (e.g., *T*, ΔBIC, or ΔAIC) differentially weighs fit and parsimony in determining model ranking (for reviews see Kuha, 2004 and Vrieze, 2012), under the same data and model conditions PAV in ranking could occur for some selection indices but not others. This is because data/model conditions rendering one selection index close to its decision threshold (implying high chance of PAV in ranking within-sample) may render another far from its decision threshold (implying low chance of PAV in ranking within-sample). For example, in a sample where ΔBIC ranges from −3 to 4 across allocations (implying PAV in ranking), ΔAIC may range from 3 to 7 (implying no PAV in ranking).

> *Hypothesis 2:* Assuming nonequivalent structural models, PAV in model ranking is more likely to arise when sampling error is higher (i.e., lower item communalities, lower *N*, and particularly their combination).

Modest *N* and low item communalities are widespread in parcel applications, and their existence is used as one motivation for parceling (e.g., Little et al., 2013; Matsunaga, 2008; Plummer, 2000; Williams & O'Boyle, 2008; Yang et al., 2010). For each model, there will be greater sampling error in selection index values in the context of lower *N* and lower item communalities[5] (see also Bandalos, 2002; MacCallum et al., 1999; Matsunaga, 2008; Meade & Kroustalis, 2006). Assuming some structural differences between models, the influence of sampling error on fit should not only be (i) allocation-specific but also (ii) model-specific (i.e., $\boldsymbol{\Delta}^{a}_{SE_p} \neq \boldsymbol{\Delta}^{b}_{SE_p}$). Regarding (i), the influence of sampling error on fit should be allocation-specific because $\boldsymbol{\Delta}_{SE_p} = \mathbf{A}\boldsymbol{\Delta}_{SE_i}\mathbf{A}'$. Regarding (ii), the influence of sampling error on fit should be model-specific because models with different structural constraints have different potential for those constraints to be inappropriate simply due to sampling error. Taken together, (i) and (ii) should allow greater potential for PAV in fit ranking proportional to the amount of sampling error. This hypothesis is illustrated in Figure 1 by contrasting the top row of plots (where sampling error is high) with the bottom row of plots (where sampling error is low). There is incidence of PAV in ranking in all four panels in the top row of plots but in only two panels in the bottom row of plots. Also, the top row of plots has on average greater magnitude of PAV in ranking.

## Simulation Study to Assess PAV in Model Ranking Within-Sample

We use a simulation study to investigate Hypotheses 1 and 2. The simulation involved a fully crossed design with 84 cells resulting from manipulating sample size (four levels), difference between structural specifications of Models *a* and *b* (seven levels), and communality size (three levels). Details of these conditions are provided later in the section titled *Manipulated design conditions,* following the presentation of the generating model. Although the simulation study pertains to a pair of models, if a researcher had multiple models to compare, results of this simulation would apply to each pair under comparison. The empirical example provided later illustrates a selection scenario with more than two models.

### Simulation Method

### Generating Model

Within each cell of the design, 500 sample datasets were generated from the following item-level model. This generating item-level model was chosen to reflect an empirical application that had employed parceling and then compared alternative structural models—Zampetakis et al. (2009).

First consider the structural portion of the item-level generating model. As diagrammed in Figure 2, there were $q = 5$ factors, $\boldsymbol{\eta}_i = [\eta_i^{pro}, \eta_i^{emo}, \eta_i^{cre}, \eta_i^{att}, \eta_i^{ent}]$, representing proactivity, emotional intelligence, creativity, attitudes toward entrepreneurship, and entrepreneurial intention, such that

$$\mathbf{B}_i = \begin{bmatrix} 0 & B_{12} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ B_{31} & B_{32} & 0 & 0 & 0 \\ B_{41} & B_{42} & B_{43} & 0 & 0 \\ B_{51} & 0 & B_{53} & B_{54} & 0 \end{bmatrix}. \quad (8)$$

Generating values of $B_{41}$ and $B_{43}$ were manipulated in the study design (see *Manipulated design conditions* section, below). On a

[5] This can be explained using notation from Appendix A, as follows. For each model, lower *N* implies that elements of $\mathbf{C}_{u\xi_i}$ and $\mathbf{C}_{\zeta u_i}$ and off-diagonal elements of $\mathbf{C}_{uu_i}$ depart from 0 by chance alone, which increases the impact of $\boldsymbol{\Delta}_{SE_i}$ in Equation (1). Also, lower item communalities imply greater impact of $\mathbf{C}_{uu_i}$ on $\boldsymbol{\Delta}_{SE_i}$ in Equation (1).

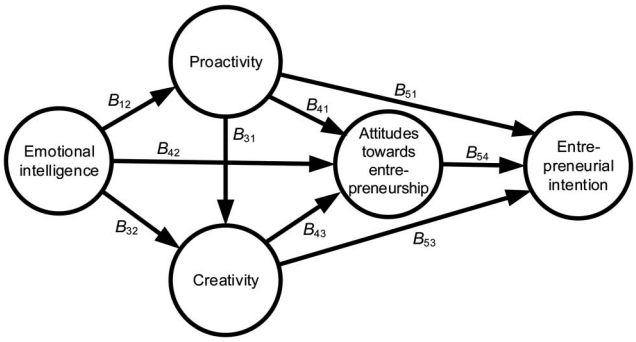


*Figure 2.* Simulation population-generating structural model. Generating values of $B_{41}$ and $B_{43}$ were manipulated in the study design; see text. The proactivity, emotional intelligence, attitudes, and entrepreneurial intention factors each had nine item indicators per factor that were subsequently parceled within factor. The creativity factor had three item indictors that were not subsequently parceled.

standardized metric, generating values of $B_{12}$, $B_{31}$, $B_{32}$, $B_{42}$, $B_{53}$, $B_{54}$ were .2 and the generating value of $B_{51}$ was .1. We specified factor residual variance values in $\mathbf{\Omega}_i$ to imply total factor variances of 1 in the population.

In the measurement portion of the item-level generating model, $\lambda_i$ and $\psi_i^2$ were manipulated in the study design (see *Manipulated design conditions* section, below). In total, there were $m = 39$ multivariate normally distributed item indicators. Specifically, there were nine item indicators per each $\eta_i^{pro}$, $\eta_i^{emo}$, $\eta_i^{att}$, and $\eta_i^{ent}$ factor. These unidimensional items were subsequently parceled within factor, as described in the next subsection. Additionally, we generated three item indictors for $\eta_i^{cre}$ which were not subsequently parceled. The latter was done because, in our literature review of applications employing parceling and model selection, sometimes researchers had several factors with parcel-indicators but one or more factors with item-indicators. M*plus* 7.3 (Muthén & Muthén, 1998-2014) was used for data generation.

## Parceling

For each of the item-level sample datasets generated per cell of the simulation design, 100 item-to-parcel allocations were randomly generated. The fact that allocations were generated randomly does not limit the relevancy of the simulation with respect to purposive allocations. This is because different kinds of purposive allocations (i.e., different purposive parceling strategies) can arise by chance as special cases of random allocations. Specifically, in a given sample, for each of the $\eta_i^{pro}$, $\eta_i^{emo}$, $\eta_i^{att}$, and $\eta_i^{ent}$ factors, we randomly assigned the nine items per factor to three parcels per factor. Each parcel score was obtained by averaging the items allocated to that parcel. This yielded 50,000 parcel-level datasets per cell. Across all 84 cells, there were thus 4,200,000 parcel-level datasets. SAS 9.4 and R were used for parceling and data management.

## Fitted Models *a* and *b*

Each parcel-level dataset was fit with two parcel-level models, *a* and *b*, using maximum likelihood estimation. M*plus* 7.3 was used for model fitting. Models *a* and *b* only differed in the structural specification.

In the model selection literature, it is considered realistic for none of the models under comparison to perfectly structurally match the generating model (for review see Preacher & Merkle, 2012). Hence, in our simulation the structural specification of *both* Models *a* and *b* were simpler than the generating model. This is consistent with our theoretical framework, which allows for structural model error. Specifically, fitted Models *a* and *b* both fixed $B_{51} = 0$, unlike the generating model in Equation (8). Whereas other kinds of structural error could be introduced,[6] its nature and presence is not central to the testing of our hypotheses. Model *a* differed from *b* in that Model *a* imposed the constraints $B_{41} = 0$ and $B_{43} = 0$, whereas Model *b* did not. As such, $df = 2$ when comparing Model *a* versus *b*. Model *a* is nested in *b*. This model comparison is similar to that employed in applications we reviewed that used parceling and then compared models that differed in structural specifications. Note that there is nothing about the phenomenon of PAV in model ranking that is limited to nested models.[7] We chose nested models for illustration so that we could include the LRT among the selection indices considered.

In the fitted model, the factor residual variances were constrained in a model-based way such that all total factor variances were 1 (e.g., Steiger, 2002). Thus, all estimated structural path coefficients are interpretable as standardized effects.

For each of the 50,000 parcel-level datasets per cell, we recorded the maximized likelihoods for Models *a* and *b*, denoted $L^a$ and $L^b$. These were used to calculate three statistics useful in model selection: the LRT statistic, $T = -2[\ln L^a - \ln L^b]$, $\Delta\text{BIC} = -2[\ln L^a - \ln L^b] + \ln N(k^a - k^b)$, and $\Delta\text{AIC} = -2[\ln L^a - \ln L^b] + 2(k^a - k^b)$. Support for Model *b* over *a* corresponds with $T(df) > T_{crit}(df)$, and also with $\Delta\text{BIC} > 0$ and with $\Delta\text{AIC} > 0$.

## Manipulated Design Conditions

**Sample size.** The 4 sample sizes used were $N = 150$, 250, 350, and 450. These modest to large sample sizes are larger than most of the *N*s considered in previous work on PAV to demonstrate that, importantly, PAV is not confined to low *N*s. In contrast, $N = 75$–250 were used in Sterba (2011) and Sterba and MacCallum (2010). The *lowest N* used in the present design ($N = 150$) is close to the average *N* used in structural equation modeling (SEM) applications, according to Baumgartner and Hornburg's (1996) review.

**Communality size.** The three item communality sizes used were: low = .16 (corresponding with $\lambda_i = .4, \psi_i^2 = .84$), medium = .30 (corresponding with $\lambda_i = .55, \psi_i^2 = .6975$), and high = .49 (corresponding with $\lambda_i = .7, \psi_i^2 = .51$). Recall that the *i* subscript indicates item level. These communality conditions were chosen using the Spearman-Brown prophecy formula to imply a particular scale reliability for each parceled factor (as in Sterba, 2011 and Sterba & MacCallum, 2010). When nine items/factor

[6] Note that Cudeck and Browne's (1992) method for introducing model error could not be used here because our interest was in introducing model error into the structural model specifically.

[7] PAV in model ranking using ΔBIC or ΔAIC can also occur with nonnested models (including models with the same degrees of freedom); this was confirmed in additional pilot simulations (not shown).

were allocated to three parcels/factor, implied scale reliability was, according to Nunnally and Bernstein (1994), excellent (.90) in the high condition, above-satisfactory (.80) in the medium condition, and .63 in the low condition. These item communalities and item-to-parcel ratios imply average parcel-communalities of .74, .56, and .36 for the high, medium, and low conditions. Population error variances were chosen to make all item variances = 1.0. Note that more PAV would be anticipated when item loadings within-factor in the population are unequal (see Case II above), so our equal-loading generating model provides a conservative depiction of potential PAV.

**Structural differences between Models *a* and *b*.** In some settings, differences between models can be defined using a metric of an overall fit index or selection index. For instance, the size of the noncentrality parameter of the noncentral $\chi^2$ distribution and its associated *df* (e.g., Fan & Sivo, 2005) could be used or the differences in expected BIC (defined in Preacher & Merkle, 2012) could be used. In the present study, however, we cannot quantify structural differences between Models *a* and *b* with respect to such fit metrics. This is because, whereas our generating model is an item-level model, fitted models *a* and *b* are parcel-level models that do not have one true fit difference. Rather, they have a distribution of fit differences across possible allocations.

In the present study, Models *a* and *b* differ only in the structural model; thus, we can instead define the difference between Models *a* and *b* parametrically, with respect to the size of the standardized structural coefficients in $\mathbf{B}_i$ that are constrained to 0 in Model *a* but freely estimated in Model *b*. Recall that standardized coefficients $B_{41}$ and $B_{43}$ were fixed = 0 in fitted Model *a* but freely estimated in fitted Model *b*. In the generating model in Equation (8), $B_{41}$ and $B_{43}$ refer to effects of proactivity and creativity on attitudes. The standardized coefficients $B_{41}$ and $B_{43}$ both have seven different generating values in the simulation design: 0, .05, .10, .15, .20, .25, .30. The six nonzero values of $B_{41}$ and $B_{43}$, ranging from .05 to .30, were chosen so that the variance in the latent attitudes factor *jointly* explained by proactivity and creativity factors ranges from small $\Delta R^2 = .01$ to large $\Delta R^2 = .24$ (Cohen, 1988). For instance, a small $\Delta R^2$ of .01 for the joint contribution of proactivity and creativity corresponds with $B_{41} = B_{43} = .05$ and a large $\Delta R^2$ of .24 corresponds with $B_{41} = B_{43} = .30$. Even for the seventh generating value, when both $B_{41}$ and $B_{43} = 0$ (and thus $\Delta R^2 = 0$), Models *a* and *b* are still structurally different (i.e., they are not equivalent models). Hence, there could still be some PAV in model ranking under this condition, because both coefficient values will not be exactly 0 in the sample.

## Simulation Results and Discussion

**Converged proper solutions.** Between 96.3% and 100% (average 99.7%) of the 50,000 allocations × samples per cell yielded parcel-solutions that converged for Models *a* and *b*. Between 86.0% and 100% (average 98.5%) yielded parcel-solutions that were both converged and proper for Models *a* and *b*. Results from the latter solutions are summarized in subsequent sections. The probability of a solution being converged or proper was virtually uncorrelated ($r = .08$ and $r = -.01$, respectively) with the size of structural differences between models, but was nonlinearly related to communality and sample size. Specifically, nonconvergence and improper solutions arose largely in the cells with both low-communality and $N = 150$; in these cells, 96–98% of parcel-solutions converged and 86–87% were converged and proper. For all other communality and sample size combinations, >99% of parcel-solutions converged and >97% were converged and proper.

**PAV in model ranking.** To address Hypotheses 1 and 2, we calculated the model ranking for each of the repeated parcel-allocations in each sample per cell. For the LRT, ranking of Model *b* over *a* (i.e., support for *b*) corresponded to a *p* value <.05. Ranking of *b* over *a* also corresponded with $\Delta BIC > 0$ or $\Delta AIC > 0$. Recall that *PAV in model ranking within-sample* is defined as when ranking preference flips from *a* to *b* or *b* to *a* in a nonzero proportion of allocations within-sample. If there is PAV in model ranking within-sample, the ranking obtained in empirical practice could be contingent on the particular item-to-parcel allocation chosen. First, we focus on results involving the *incidence* of PAV in model ranking within sample. Later, we focus on results involving the *magnitude* of PAV in ranking, among samples where PAV arose.

There are alternative ways to display results involving the *incidence of PAV in ranking*. We adopt two possible ways of depicting incidence—the proportion of samples where ≥1% of allocations exhibit PAV in model ranking, and also the proportion of samples where ≥5%, of allocations exhibit PAV in model ranking. Results pertaining to the incidence of ≥1% PAV in ranking are provided in the Online Appendix. Results pertaining to the incidence of ≥5% PAV in ranking show the same pattern but with a reduction in rates; they are provided here in Figures 3, 4, and 5. In particular, the *y*-axis of Figures 3, 4, and 5 is the proportion of samples where ≥5% of allocations exhibit a flip in model ranking—meaning that they prefer a different model than the majority of the allocations within-sample. The *x*-axis of Figures 3, 4, and 5 is the average selection index value per cell.

In Figure 3, the *x*-axis is the average LRT statistic value per cell. Each panel in Figure 3 corresponds with a different sample size. Within each panel, each curve represents a different communality size: top curve = low communalities; middle curve = medium communalities; bottom curve = high communalities. In Figure 3, each of the seven dots represents a particular size of structural differences between Models *a* versus *b*. Dots are connected by splines. The dot size is proportional to the size of the structural difference between Models *a* and *b* (i.e., with the largest dot representing $\Delta R^2 = .24$). In Figure 3, the gray vertical line represents the decision threshold, that is, the critical value of $T_{crit}(df = 2) = 5.991$. An obtained LRT statistic exceeding this critical value indicates preference for Model *b;* otherwise Model *a* is preferred.

In support of Hypothesis 1, Figure 3 shows that incidence of PAV in model ranking peaks as the obtained LRT statistic approaches the decision threshold of $T_{crit}$. Importantly, notice that the nature of the peak is not a sharp narrow spike in the immediate vicinity of the decision threshold; rather, we see a gradually increasing and then decreasing amount of PAV spanning a broad range of LRT statistic values both before and after the decision threshold (e.g., from $T = 3$ to $T = 20$). Also note that, for this model pair, PAV peaks at medium-sized structural differences when *N* is smaller (i.e., dots are medium sized at peak of curve when $N = 150$). When *N* is larger, however, PAV peaks at smaller-sized structural differences (i.e., dots are smaller at peak of curve when $N = 450$).

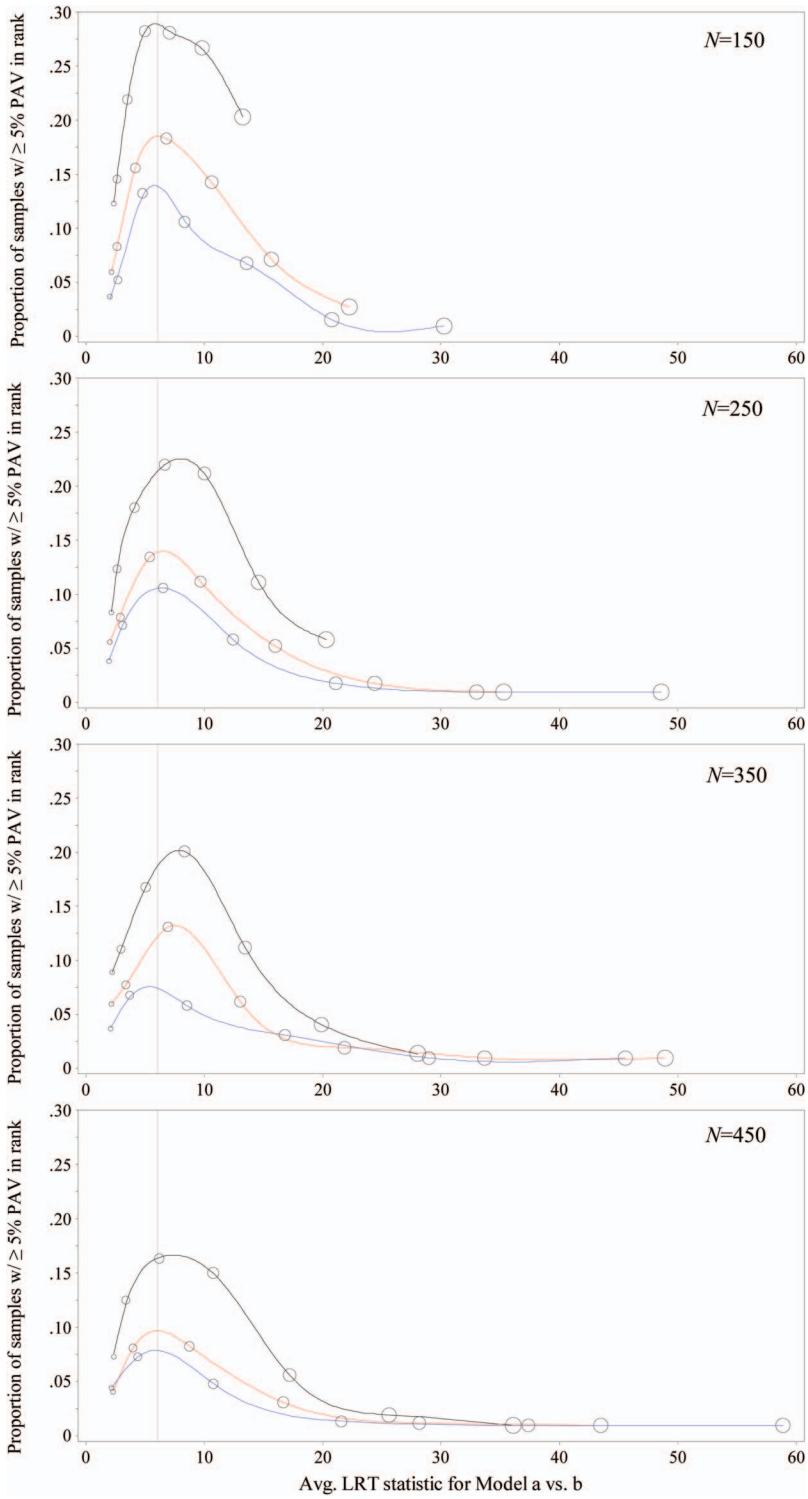


*Figure 3.* Proportion of samples per cell with ≥5% parcel-allocation variability (PAV) in model ranking versus average likelihood ratio test (LRT) statistic per cell. In each panel, top curve = low communalities; middle curve = medium communalities; bottom curve = high communalities. The seven dots per curve are connected by a spline and represent 7 sizes of structural differences between Models *a* and *b,* ranging from small to large. Vertical bar = decision threshold (i.e. critical value of 5.99) for LRT. See the online article for the color version of this figure.

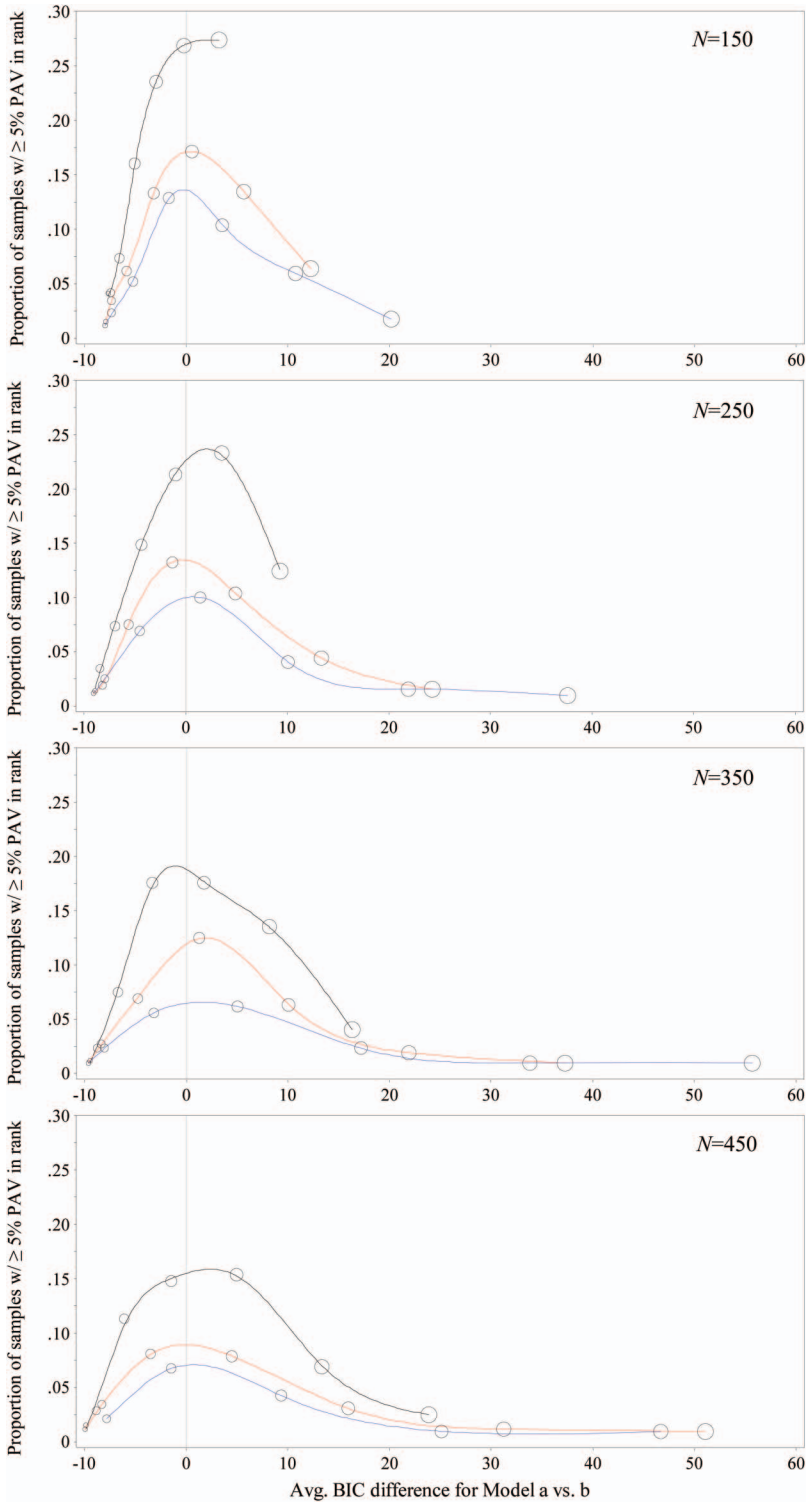


*Figure 4.* Proportion of samples per cell with ≥5% parcel-allocation variability (PAV) in model ranking versus average change in Bayesian information criterion (ΔBIC) per cell. In each panel, top curve = low communalities; middle curve = medium communalities; bottom curve = high communalities. The seven dots per curve are connected by a spline and represent seven sizes of structural differences between Models *a* and *b,* ranging from small to large. Vertical bar = decision threshold for ΔBIC (i.e., 0). See the online article for the color version of this figure.

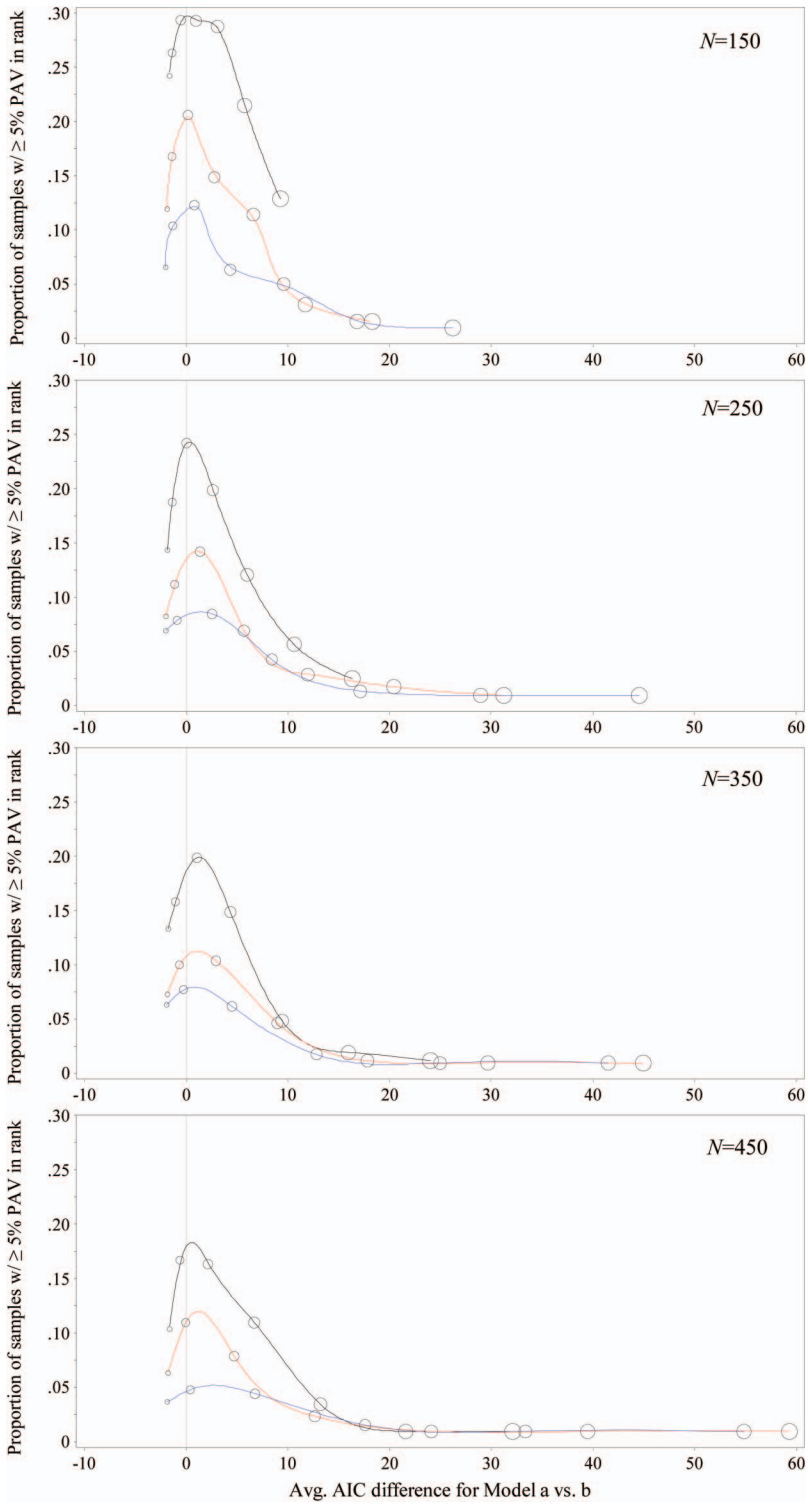


*Figure 5.* Proportion of samples per cell with ≥5% parcel-allocation variability (PAV) in model ranking versus average change in Akaike information criterion (ΔAIC) per cell. In each panel, top curve = low communalities; middle curve = medium communalities; bottom curve = high communalities. The seven dots per curve are connected by a spline and represent seven sizes of structural differences between Models *a* and *b*, ranging from small to large. Vertical bar = decision threshold for ΔAIC (i.e., 0). See the online article for the color version of this figure.

In support of Hypothesis 2, peak frequency of PAV occurrence is higher when $N$ and/or item communalities are lower. For example, when both factors are low, PAV in ranking occurs in up to 28% of samples. Importantly, when only one of these factors is low, the risk of PAV in model ranking remains, in line with Hypothesis 2. That is, for lower sample size ($N = 150$)—but medium to high communalities—PAV in model ranking can still occur in 13–18% of samples. Likewise for low communalities—but medium to high $N$—PAV in model ranking can still occur in 17–20% of samples. Note that even combinations of medium to large $N$ and medium to large communalities still can give rise to PAV in model ranking in up to 10–12% of samples.

In Figure 4, the $x$-axis is average ΔBIC per cell and in Figure 5 the $x$-axis is average ΔAIC per cell. In Figures 4 and 5, the gray vertical line represents the decision threshold of 0. Supporting *Hypothesis 1*, in Figures 4 and 5 the incidence of PAV in model ranking again peaks when the selection index is closer to its decision threshold (0). Notably, there was elevated PAV across a wide range of index values surrounding the decision threshold (e.g., ΔBIC of −9 to 15). Only when the absolute values of cell-average ΔBIC or ΔAIC were extremely large was there no incidence of PAV in model ranking (i.e., Model $a$ or Model $b$ was always preferred). Additionally, supporting *Hypothesis 2,* in Figures 4 and 5 the peak in PAV for model ranking is higher when communalities are lower and/or $N$ is lower.

It is important to note that the data/model conditions corresponding to peak risk of PAV in model ranking differ somewhat from one selection index to another across Figures 3, 4, and 5. For instance, for this pair of models, PAV incidence peaks at larger structural differences (i.e., larger dot sizes) when using ΔBIC compared with LRT or ΔAIC, and when using LRT compared with ΔAIC. This is because different sizes of structural differences are needed for each index to be close to its decision threshold. The LRT statistic will equal its decision threshold when the fit difference $-2[\ln L^a - \ln L^b]$ equals $T_{crit}(df)$; ΔBIC will equal its decision threshold when the fit difference equals its penalty term $\ln N(k^a - k^b)$; ΔAIC will equal its decision threshold when the fit difference equals its penalty term $2(k^a - k^b)$. Consider, for instance, the top curve in the top panel of Figures 3, 4, and 5 (i.e., lower communalities and $N = 150$). A larger fit difference, and thus a larger structural model difference (i.e., larger dot size), is needed for ΔBIC to equal its decision threshold (10.02) than for LRT to equal its decision threshold (5.99) or ΔAIC to equal its decision threshold (4.0). This pattern reflects the fact that Model $b$ is the more complex model and ΔAIC generally has a lower bar for preferring the more complex model than ΔBIC (Kuha, 2004; Preacher, Zhang, Kim, & Mels, 2013). What this means in the context of PAV in model ranking is that there is a wide range of structural differences under which PAV could occur for at least one selection index. That is, given a particular $N$ and communality size, for this model pair there could be an elevated risk of PAV using ΔAIC for a comparison involving small structural differences, but elevated risk of PAV using LRT or ΔBIC for a comparison involving medium to large structural differences.

We have thus far considered the incidence of PAV in ranking. Next, for samples where PAV in model ranking arises in a nonzero proportion of allocations within-sample, we also consider the magnitude of PAV in ranking within-sample. We quantify the *magnitude of PAV in ranking* as the proportion of allocations per sample preferring the least favored model. Hence, the maximum magnitude boundary is .50. In Figure 6, the magnitude of PAV in ranking in a sample (on the $y$-axis) is plotted against the average selection index value (LRT statistic or ΔBIC, or ΔAIC) in that sample (on the $x$-axis). The three rows of Figure 6 depict this relationship for low, medium, and high communalities. In the online appendix (see Supplemental Materials), a parallel plot is provided depicting this relationship for $N = 150$, 250, 350, and 450 separately; the same pattern of results in Figure 6 across low to high communalities is found in the online appendix across low to high $N$. Each data point in Figure 6 is a sample where PAV in ranking occurred. Figure 6 shows that the magnitude of PAV in ranking within-sample peaks (with half of allocations preferring Model $a$ and half preferring Model $b$) when the average selection index value within-sample is closest to the selection index's decision threshold. The decision threshold is represented by a vertical bar. Additionally, note how the white space around the decision threshold decreases as sampling error decreases. This implies that samples with average selection index values near the decision threshold tend to exhibit peak PAV magnitude (i.e., near .50) when sampling error is high (top row), but tend to exhibit a whole range of PAV magnitudes when sampling error is low (bottom row). This is because there is more across-allocation variability in index values when sampling error is higher. Thus, in samples with average index values near the decision threshold, a larger proportion of allocations within-sample tend to span the decision threshold when sampling error is higher than when it is lower. Consequently, for a single sample with an average selection index value close to the decision threshold, it is easier to predict the magnitude of PAV in ranking when sampling error is high.

**Summary.** In sum, there was elevated potential for PAV in model ranking across a range of selection index values around each index's decision threshold. Within this range, the incidence and magnitude of PAV in model ranking depended on sampling error (i.e., $N$ and item communalities). It is possible to have PAV in ranking for one selection index but not other indices. Eliminating risk of PAV in model ranking required structural differences rendering each index far from its decision threshold (e.g., very inappropriate constraints imposed in model $a$ but not $b$ in this simulation) plus either medium-to-high communalities or $N \geq 250$.

## Model Ranking in the Item-Solution Versus Modal Parcel-Solution

When PAV in model ranking exists within-sample, a researcher is confronted with the question of which model to select. If the researcher picks only one allocation, the model ranking may differ if another single allocation were chosen. Instead, one suggestion is to conclude in favor of the model selected in the highest proportion of allocations within sample. This is here termed the *across-allocation modal ranking* (AMR). To implement this suggestion, first the within-sample proportion of allocations (WPA) preferring Model $b$ is obtained. If this WPA $>$ .50, the AMR would select

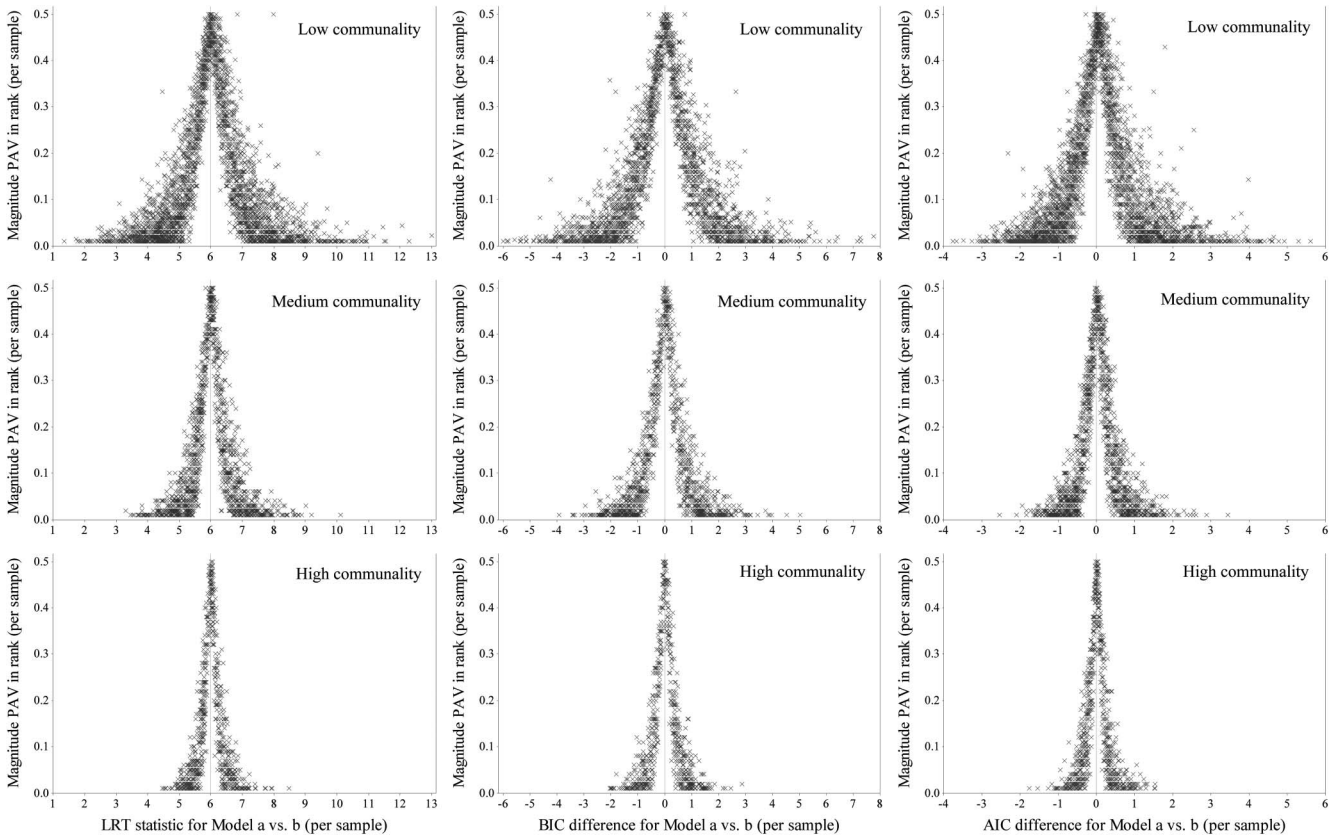


*Figure 6.* Magnitude of parcel-allocation variability (PAV) in model ranking. Each X = a sample where PAV in model ranking occurred. *Magnitude* of PAV in rank = proportion of allocations/sample preferring the least-favorable model; maximum magnitude = .50. Vertical bar = decision threshold. LRT = likelihood ratio test; BIC = Bayesian information criterion; AIC = Akaike information criterion.

Model *b*; otherwise, *a* would be selected.[8] The suggestion to use AMR reflects the *principle of aggregation* (discussed in other settings by Little et al., 2013; Matsunaga, 2008; Nunnally, 1978; Rushton, Brainerd, & Pressley, 1983) in that any one parcel-allocation is less representative than the average of many such parcel-allocations from the same parcel-allocation distribution. Additional rationale for this suggestion stems from the finding that the AMR, on average, matches the model ranking obtained if *item-level* versions of Models *a* and *b* were fit, instead of parcel-level Models *a* and *b*. This result is useful to note because researchers often are initially interested in an item-level solution but end up using parcel-level solutions as a proxy for practical reasons, as described in the introduction.

Specifically, the item-solution ranking matches the AMR in 98% of samples for LRT, 99% of samples for ΔBIC, and 98% of samples for ΔAIC.[9] The accuracy of this within-sample match varied slightly from cell to cell, as shown in Figure 7, right column. Each datapoint in the right column of Figure 7 represents a design cell. The *y*-axis depicts the proportion of samples per cell where the parcel-solution AMR matches the item-solution ranking. Note the concave-up quadratic relationship across cells in Figure 7, right column. The minimum of this curve corresponds to cells with the worst match between item solution ranking versus parcel-solution AMR (i.e., minimum 93% match for LRT, 92% for ΔBIC, and 93% for ΔAIC). This minimum occurs, on the *x*-axis, where cell-average WPA is around .50. Additionally, as shown in Figure 7, left column, the proportion of item solutions that prefer Model *b* in each cell (*y*-axis) is correlated >.999 with the cell-average WPA (*x*-axis). Each datapoint in Figure 7, left column, again represents a design cell.

Taken together, Figure 7 indicates that, if one has decided to parcel, selecting the model which is preferred in the highest proportion of parcel-allocations within sample can be justified on the grounds that this same ranking would be expected from an item-level analysis, on average. This result is noteworthy because it differs from the relationship between *absolute model fit* of item-solutions and parcel-solutions for a single model in isolation (see Bandalos, 2002; Landis et al., 2000; Meade & Kroustalis,

[8] If WPA is exactly .50 in practice, we would suggest substantially increasing the number of random allocations within sample and then reporting the WPA as well as the AMR (see Discussion).

[9] Compared to parcel-solutions, a similar proportion of item solutions were converged and proper (range across cells: 83.6%–100%; average: 98.4%). The Figure 7, left column, results were re-run including vs. excluding samples where item-solutions were nonconverged and/or improper but some parcel-solutions were converged and proper. The pattern of results was the same.

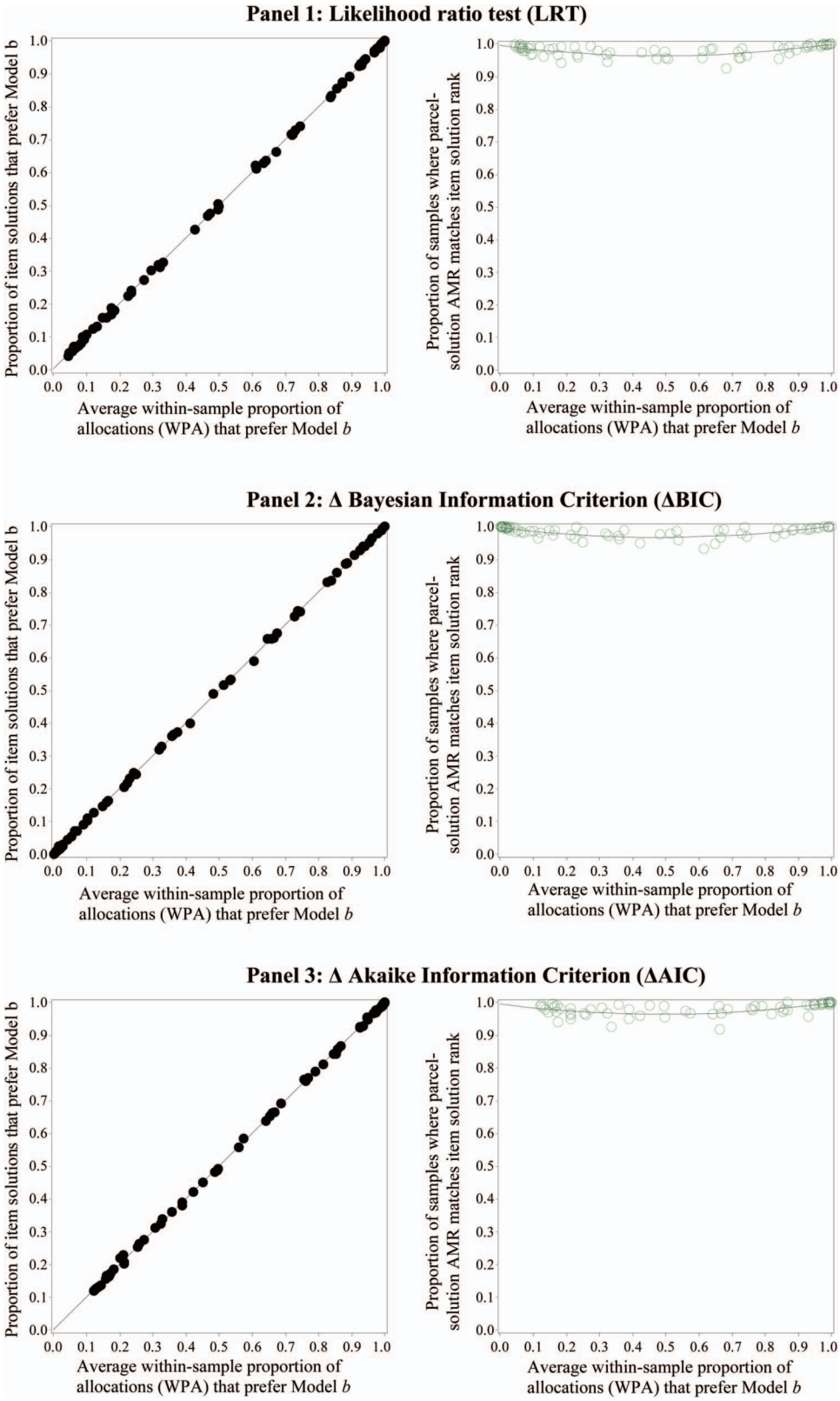


*Figure 7.* Comparing item-solution model ranking versus parcel-solution across-allocation modal ranking (AMR). Each data point represents a design cell. See the online article for the color version of this figure.

2006; Nasser & Takahashi, 2003; Nasser & Wisenbaker, 2003; Rogers & Schmitt, 2004). For a single model in isolation, the absolute fit of an item solution tends to be systematically worse than the across-allocations average fit of a parcel-solution. In the context of evaluating absolute fit, the *df* of the item-level model greatly exceeds the *df* of a parcel-level model; the larger dimension of the item-solution's error covariance matrix, for instance, provides systematically more potential for unmodeled error covariances to cause misfit (see Bandalos, 2002; Sterba, 2011). In the context of model selection among structural models, the pair of item-level Models *a* and *b* and the pair of parcel-level Models *a* and *b* have the same $\Delta df$ and impose/relax the same structural constraints. We have shown here that ranking among structural models need *not* be systematically different between the item-solution ranking and the parcel-solution AMR.

## Software Tools for Assessing PAV in Model Ranking Within-Sample

The above simulation study was performed under the conservative situation of no measurement-model-specific model error (consistent with the assumption of most researchers using parceling) and equal item-loadings within-factor. Even under these stringent conditions, simulation results showed PAV in model ranking occurring under many different conditions depending on the combination of sample size, communality size, structural difference between models, and selection index used. The context of real-world data may present different—and potentially less ideal—conditions such as smaller sample sizes and measurement-model-specific model error (e.g., induced by using categorical rather than metric items). Hence, in practice it is useful and diagnostic for researchers to assess the degree of PAV in model selection in their own sample, for their own models of interest. To that end, we supply a software utility to allow researchers to do so. Previously, Sterba and MacCallum (2010) produced a SAS-based software utility to allow researchers to repeatedly, randomly allocate items to parcels for a single SEM model in isolation, where model-fitting occurred in a SEM package of choice. Quick and Schoemann (2012) converted this utility from SAS into R, again for a single model in isolation. The present R tool (*PAVranking*) extends that of Quick and Schoemann (2012) to the context of model selection between competing SEM models, generalizes it to handle nonconverged and improper solutions, and supplies a variety of new output relevant to assessing PAV in model ranking, as described below.

To use the *PAVranking* R utility,[10] the researcher provides an item-level dataset and specifies (in *lavaan* input format) two competing SEM models of interest at a time that differ in the structural model specification. The researcher also specifies the desired numbers of: parcels per factor, items per parcel, and random item-to-parcel allocations. Both models are fit repeatedly (using *lavaan,* Rosseel, 2012) to that number of randomly generated allocations. In addition to providing output information relevant to each model considered separately (the across-allocations average, standard deviation, and range of parameter estimates, standard errors, and absolute fit indices; Quick & Schoemann, 2012; Sterba & MacCallum, 2010), the program provides the following output relevant to model selection: the proportion of allocations in which the LRT is significant, the proportion of allocations where each model is selected according to BIC and AIC, the average size of the $\Delta$BIC and $\Delta$AIC, and the across-allocations average, standard deviation, and range of the LRT. Furthermore, plots are also automatically generated for the distribution of $\Delta$BIC, $\Delta$AIC, and LRT *p* values across allocations within-sample. Additionally, the researcher can specify whether he or she is comparing multiple models, so that the same set of random item-to-parcel allocations can be used for each pairwise model comparison. In the empirical example section below, we use this software to quantify PAV in model ranking and aid in drawing substantive conclusions.

## Empirical Example

Here we consider an empirical analysis where the goal is model selection. We use this example to demonstrate interpretation and reporting of selection results in the context of PAV. This example is pedagogically illustrative because it includes commonly encountered model specifications. Furthermore, this example generalizes the context of the simulation study in that it involves multiple model comparisons, a different *N,* and categorical rather than continuous items. The simulation results are relevant to each model comparison pair.

The empirical example involves four latent factors identified and used in previous research (e.g., Hill, Payne, Jackson, Stine-Morrow, & Roberts, 2014; Smith, 2013). Each was measured on $N = 102$ undergraduates in the 1988 Computer-Assisted Study (Latane, 1989). The first latent factor was belonging social support (i.e., perceived availability of people to do things with) measured with 12 binary items (Cohen & Hoberman, 1983, Interpersonal Support Evaluation list). The next three factors were openness to feelings, general anxiety, and perceived vulnerability to stress, each measured with 8 5-point ordinal items (from the Neuroticism-Extraversion-Openness [NEO] personality inventory, Costa & McCrae, 1985). Items were randomly allocated to parcels 100 times in the following manner. The 12 items on the first factor were parceled into three parcels of four items each (as in, e.g., Brookings & Bolton, 1988). The eight items on each of the next three factors were parceled into three parcels of three, three, and two items per factor. The same 100 item-to-parcel allocations were used in all fitted models.

In this example, research interest focused on four parcel-level models that involve predicting anxiety. These four models differ only in the structural model specification (see Figure 8). The parcel-level measurement model specification was the same in all fitted models (as in the simulation study). As depicted in Figure 8, in Model 4, anxiety is predicted by belonging support, stress vulnerability, and openness to feelings. Models 1, 2, and 3, also depicted in Figure 8, are simpler competing structural models. In Model 1, openness to feelings predicts both social support and vulnerability (Eldesouky, 2012), which both, in turn, predict general anxiety (Thoits, 1984). In other words, emotional awareness may both elicit more social support (which may reduce anxiety) and increase resilience to stress (which also may reduce anxiety).

[10] This utility will be made available online at www.vanderbilt.edu/psychological_sciences/bio/jason-rights and www.vanderbilt.edu/peabody/sterba/appxs.htm and in the semTools R package https://cran.r-project.org/web/packages/semTools/index.html.

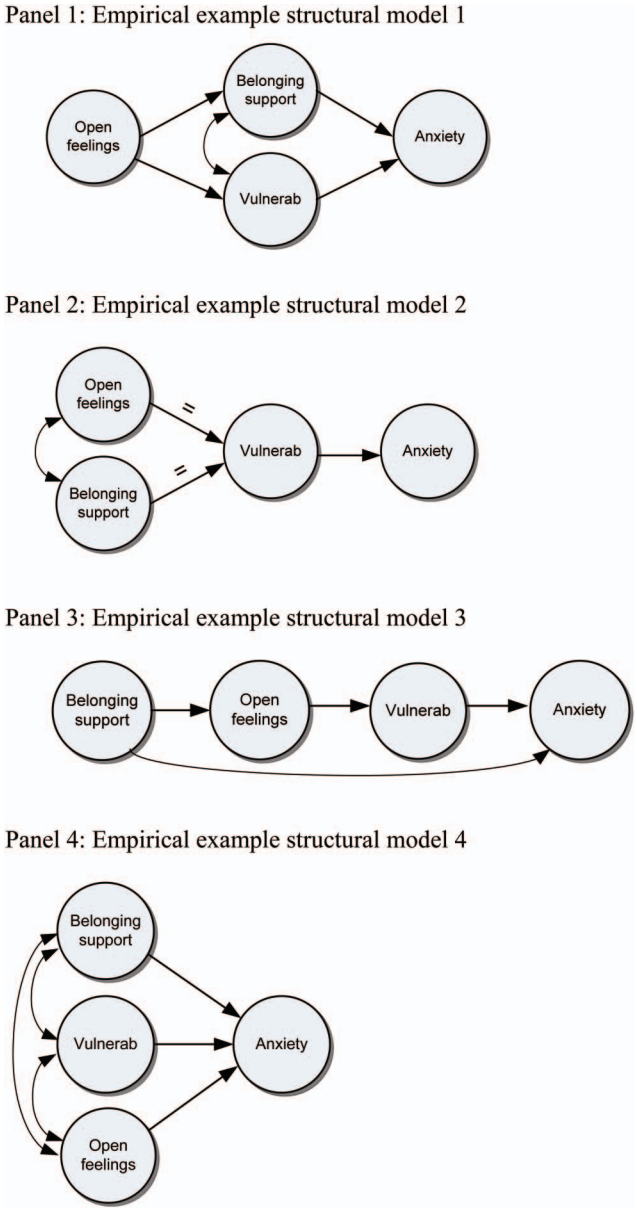


*Figure 8.* Empirical example structural models. Measurement models for all latent factors shown involve three parcel indicators per factor. In Panel 2, model-based standardization was used for variables involved in the equality constraint. See the online article for the color version of this figure.

In Model 2, openness and belonging support have equal effects on stress vulnerability, which then predicts anxiety (Gershuny & Sher, 1998; Sexton, Norton, Walker, & Norton, 2003; Suarez, Bennett, Goldstein, & Barkow, 2008; Williams, Rau, Cribbet, & Gunn, 2009). Finally, in Model 3, the amount of social support predicts anxiety level directly (e.g., Karevold, Røysamb, Ystrom, & Mathiesen, 2009), but also through affecting openness and vulnerability. We were substantively interested in comparing Models 1, 2, and 3 each to Model 4. Of course more (or all) model comparisons would be possible. Models 1, 2, and 3 were each nested in Model 4 so we elected to use LRT as well as ΔBIC and ΔAIC selection indices for these comparisons, as follows.

First consider the Model 1 versus 4 comparison. As shown in Table 1, there was PAV in model ranking for all three selection indices, and on average (across parcel-allocations within-sample), all three selection indices supported Model 4. Specifically, 93–98% of allocations preferred Model 4, depending on the selection index. For the Model 2 versus 4 comparison, in Table 1, there was PAV in model ranking for two out of three selection indices (ΔBIC and LRT). Specifically, for ΔBIC there was support for Model 4 in 77% of allocations and for LRT there was support for Model 4 in 98% of allocations. On average across allocations all three selection indices supported Model 4. Finally, for the Model 3 versus 4 comparison, in Table 1, there was no PAV in model ranking. All three selection indices supported Model 3 in 100% of allocations. These empirical results are consistent with our earlier simulation results in that they illustrate the possibility of PAV occurring in some model selection index values but not others, within sample.

Taken together, these empirical results indicate support for Model 4 over 1 or 2 and support for Model 3 over 4. Compared to Model 4, Model 3 provides a more parsimonious representation of a pathway to anxiety. We report AMR results because they are typical of the distribution of parcel-allocations in this sample. If we had selected and reported results from only one parcel-allocation, instead of aggregate results across a distribution of parcel-allocations, we could have obtained a different model selection outcome that was quite atypical among possible parcel-allocations.

Though outside of the scope of this illustration, in practice, other study designs could be considered to strengthen the grounds for causal inference, such as measuring anxiety at a later timepoint. Also, future research could further study the potential pathway to anxiety in Model 3 by, for example, experimentally manipulating the amount of social support provided.

## Overall Discussion

Applied researchers widely use parceling when their research interest lies in comparing competing structural models each specified using the same parcel-level measurement model (e.g., Booth et al., 2013; Daspit et al., 2013; Dunkley et al., 2014; Geiser et al., 2013; Hankonen et al., 2014; Flack et al., 2011; Gallagher et al., 2009; Gellert et al., 2012; Jackson & Gaertner, 2010; Kuhn & Holling, 2009; Liao et al., in press; Mairet et al., 2014; Malmberg & Little, 2007; Martin et al., 2011; Nouwen et al., 2009; Owuamalam et al., 2014; Segrin et al., 2013; Sierau & Herzberg, 2012; Winkler et al., 2015; Zampetakis et al., 2009; Zheng et al., 2014). Although much applied research employs parceling with moderate sample sizes in the context of structural model selection, we lacked understanding of the within-sample consequences of PAV in this context. The present article filled this gap.

In summary, this article first extended a theoretical framework to show how structural differences, in the absence of sampling error, can induce PAV in model selection index values when item loadings are unequal within-factor in the population. This framework also showed how, more generally, sampling error plus structural differences between models can induce PAV in model selection index values when item loadings are equal or unequal within-factor in the population. We used a simulation to test two hypotheses, informed by this framework, regarding patterns of PAV in

Table 1
*Empirical Example Model Selection Results: Across-Allocation Averages, Standard Deviations, and Ranges of Selection Index Values*

| Models | LRT statistic | | | LRT $p$ value | | | ΔBIC | | | ΔAIC | | | PAV in model ranking? |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Avg | *SD* | Range | Avg | *SD* | Range | Avg | *SD* | Range | Avg | *SD* | Range | |
| 1 vs. 4 ($df$ = 1) | 13.07 | 6.69 | (.71, 29.61) | .013 | .053 | (0, .400) | 8.51 | 6.69 | (−3.85, 25.05) | 11.07 | 6.69 | (−1.29, 27.61) | Yes: LRT, ΔBIC, ΔAIC |
| 2 vs. 4 ($df$ = 3) | 17.92 | 5.55 | (7.65, 31.81) | .004 | .010 | (0, .054) | 4.22 | 5.55 | (−6.04, 18.11) | 11.92 | 5.55 | (1.65, 25.81) | Yes: LRT, ΔBIC |
| 3 vs. 4 ($df$ = 2) | .57 | .64 | (.01, 2.98) | .782 | .193 | (.225, .997) | −8.55 | .64 | (−9.12, −6.15) | −3.42 | .64 | (−3.99, −1.01) | No |

*Note.* LRT = likelihood ratio test; BIC = Bayesian information criterion; AIC = Akaike information criterion; PAV = parcel allocation variability; Avg = average; $df$ = difference in the number of free parameters between models.

model ranking. In support of these hypotheses, simulation results showed that PAV peaked when structural model differences implied selection index values close to the index's decision threshold. The incidence and magnitude of PAV at this peak was proportional to sampling error. That is, the incidence and magnitude was higher when $N$ and item communalities were lower. Potential for PAV in ranking occurring in at least 5% of allocations/sample remained elevated in a relatively broad vicinity of each index's decision threshold (e.g., $-9 < \Delta\text{BIC} \leq 15$). In the simulation, there was a wide range of structural model differences (including small, $\Delta R^2 = .01$, and large, $\Delta R^2 = .24$) across which at least one selection index was close to its decision threshold—and, thus, where there was an elevated chance of PAV in ranking. Furthermore, the potential for PAV in model ranking cannot simply be quantified as a function of data and model conditions, because it depends to some extent on the selection index used. Taken together, ensuring the absence of within-sample PAV in model ranking for all selection indices necessitated structural differences that render each index far from its decision threshold, (e.g., very inappropriate constraints in Model $a$ but not $b$), as well as either medium/high communalities or $N \geq 250$.

In practice, we suggested that researchers could report the across-allocation modal ranking (AMR) for parcel-solutions. It was shown via simulation to correspond, on average, with the item-solution model ranking. We provided software that can be used for detecting and describing PAV in model ranking, and we demonstrated its use in the context of an empirical example comparing competing structural models predicting a latent anxiety factor. Note that, in practice, there is no need to make an arbitrary cut-off designation regarding what amount of PAV in ranking within-sample is meaningful (e.g., ≥1%, ≥5%); the AMR can be reported regardless.

In addition to reporting the AMR, researchers can report the magnitude of PAV in ranking within their sample, as was done in the empirical example. Particularly if a researcher had interest in precisely estimating the magnitude of PAV in ranking, and/or if the magnitude was near .50, the researcher would want to ensure that the percent of allocations preferring a given model remained stable when the total number of allocations (here, 100) was appreciably increased (see also Sterba & Rights, in press).

## Parcel-Allocation Variability in Absolute Model Fit Versus in Model Fit Ranking

This study's findings, in conjunction with previous findings, yield the following important implications about the relationship between PAV in model ranking versus PAV in absolute fit. In a sample, a researcher could find (a) PAV in model ranking but not PAV in absolute fit of a given model, (b) PAV in absolute fit of a given model but not model ranking, or (c) PAV in both. Examples of each possibility can be inferred by combining the present results with those of Sterba and MacCallum (2010), under conditions examined in these studies. Specifically, a researcher could often encounter (a) if sampling error were low and at least one model selection index value was near its decision threshold, (b) if sampling error were high but at least one model selection index value was far from its decision threshold, and (c) if sampling error were

high and at least one model selection index value was near its decision threshold.

## Generalizability Considerations

We increased generalizability by choosing different conditions in our simulation than empirical example (e.g., presence/absence of an unparceled factor, two or more model pairs, continuous/categorical items, different *N*s). Here, we further discuss the generalizability of our simulation results in the context of using random allocations, unidimensional items per factor in the population, unequal item loadings within factor, and particular model pairs.

First, although we illustrated PAV in model selection using repeated *random* allocations (as commonly done in practice, Bandalos & Finney, 2001), we noted that this phenomenon is not limited to random allocations. If a single allocation is purposively chosen, there can still exist a hypothetical distribution of item-to-parcel allocations with the same number of items/parcel and parcels/factor.

Second, we conservatively chose to employ unidimensional items in the population in our simulation because there is the most widespread agreement on the appropriateness of parceling in this setting (e.g., Bandalos, 2002, 2008; Hagtvet & Nasser, 2004; Hall et al., 1999; Hau & Marsh, 2004; Landis et al., 2000; Little et al., 2002; Marsh & O'Neill, 1984; Marsh et al., 2013; Matsunaga, 2008; Meade & Kroustalis, 2006; Nasser-Abu & Wisenbaker, 2006; Plummer, 2000; Rogers & Schmitt, 2004; Sass & Smith, 2006; Williams & O'Boyle, 2008; Yang et al., 2010; Yuan et al., 1997). Parceling multidimensional items per factor or parceling in the context of appreciable model error in the item-level measurement model would be expected to increase the potential for PAV in model ranking, all else equal. Future research should address how measurement model misspecification (e.g., unaccounted for nonnormality, nonlinearity, or multidimensionality) could affect the correspondence between AMR parcel-solution results and item-solution ranking results described in Figure 7. It was outside the scope of this article to address how unidimensionality might be tested with the sample; there are large literatures on this topic in factor analysis, item response theory, and exploratory structural equation modeling (e.g., Bollen, 1989; Embretson & Reise, 2000; Marsh, Morin, Parker, & Kaur, 2014; Stucky et al., 2012). If a researcher intends to test and establish unidimensionality of items per factor as a prerequisite for parceling—as recommended by, for instance, Marsh et al. (2013) and Matsunaga, (2008)—this literature can be consulted for procedures.

Third, here we also employed equal-$\lambda_i$ within-factor in the simulation's item generating model. Because our Case II section showed that unequal-$\lambda_i$ within-factor can induce PAV in selection index values even in the absence of sampling error, our simulation provides a conservative lower bound for PAV that might be encountered under alternative conditions of unequal-$\lambda_i$. Future research can investigate incidence and magnitude of PAV in model ranking under broader conditions, including unequal-$\lambda_i$ within-factor.

Fourth, model pairs used in the simulation and empirical example were chosen because they mirrored those commonly employed in practice. They were used to provide proof of concept of PAV in model selection. In the simulation conditions where $B_{41}$ and $B_{43}$ equaled .05, .10, .15, .20, .25, or .30, structural differences between models were induced by placing inappropriate constraints in Model *a* that were not imposed in Model *b*. In the simulation conditions where $B_{41}$ and $B_{43}$ equaled 0, structural differences between models entailed freely estimating unnecessary/superfluous parameters in Model *b* that were fixed in Model *a*. If additional superfluous free parameters had been added to Model *b* in the latter condition, some risk of PAV in ranking would remain so long as the selection index value remained close to its decision threshold. Choosing other pairs of models in conjunction with other sample sizes will affect where ΔBIC, ΔAIC, and LRT peak in PAV relative to each other. For this reason, we suggest that researchers can use software tools to gauge the risk of PAV in ranking under their own specific model and data conditions.

## Future Research Directions

Previous studies concerning PAV in absolute fit of a single model found that it mainly occurred under high sampling error. This may have led some to assume that PAV in absolute fit of a single model is limited to this setting. However, previous studies assumed no structural model error. Using the extended theoretical framework supplied here that incorporates structural error, future research could examine the combined effects of sampling error and structural model error on PAV in absolute fit. The present findings predict a compensatory relationship where, under *low* sampling error, there could still be substantial risk of PAV in absolute fit, so long as there is modest structural model error. In other words, the present findings predict an appreciably *increased scope* of conditions evidencing PAV in absolute model fit—depending on the precise combination of structural model error and sampling error.

## Conclusions and Recommendations

Historically, parceling has been motivated as a way to combat various kinds of data suboptimalities, such as nonnormal or coarsely categorized items, when using normal-theory estimation. There are alternative ways to address some problems for which parceling had been suggested as a solution, while still fitting item-level models. For instance, advances in nonnormality-robust estimation and categorical variable estimation reduce the need for parceling in this context (e.g., Bandalos, 2008, 2014; DiStefano & Morgan, 2014). In other situations—such as the combination of complex models with moderate sample sizes—item-level models may become infeasible and parcel-level models are still considered when interest lies in structural relationships (e.g., Bagozzi & Edwards, 1998; Hau & Marsh, 2004; Little et al., 2013; Marsh et al., 2013; Matsunaga, 2008; Meade & Kroustalis, 2006; Nasser & Wisenbaker, 2003; Sass & Smith, 2006; Williams & O'Boyle, 2008; Yang et al., 2010). In this exact setting, and despite having unidimensional items in the population, we showed that PAV in model ranking arises under a variety of sample sizes, communality sizes, and differences between structural models. Researchers are encouraged to investigate the presence of PAV in model ranking within their own sample and report the AMR and WPA. This sensitivity analysis can even aid in interpreting the model ranking from a single substantively chosen allocation.

## References


Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest, Hungary: Akademia Kiado.

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

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

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

Bandalos, D. (2014). Relative performance of categorical diagonally weighted least squares and robust maximum likelihood estimation. *Structural Equation Modeling, 21,* 102–116. http://dx.doi.org/10.1080/10705511.2014.859510

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

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. http://dx.doi.org/10.1016/0167-8116(95)00038-0

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley. http://dx.doi.org/10.1002/9781118619179

Booth, T., Murray, A., Marples, K., & Batey, M. (2013). What role does neuroticism play in the association between negative job characteristics and anxiety and depression? *Personality and Individual Differences, 55,* 422–427. http://dx.doi.org/10.1016/j.paid.2013.04.001

Brookings, J. B., & Bolton, B. (1988). Confirmatory factor analysis of the interpersonal support evaluation list. *American Journal of Community Psychology, 16,* 137–147. http://dx.doi.org/10.1007/BF00906076

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33,* 261–304. http://dx.doi.org/10.1177/0049124104268644

Coffman, D. L., & MacCallum, R. C. (2005). Using parcels to convert path analysis models into latent variable models. *Multivariate Behavioral Research, 40,* 235–259. http://dx.doi.org/10.1207/s15327906mbr4002_4

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

Cohen, S., & Hoberman, H. M. (1983). Positive events and social supports as buffers of life change stress. *Journal of Applied Social Psychology, 13,* 99–125. http://dx.doi.org/10.1111/j.1559-1816.1983.tb02325.x

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

Cudeck, R., & Browne, M. W. (1992). Constructing a covariance matrix that yields a specified minimizer and a specified minimum discrepancy function value. *Psychometrika, 57,* 357–369. http://dx.doi.org/10.1007/BF02295424

Daspit, J., Tillman, C., Boyd, N., & McKee, V. (2013). Cross-functional team effectiveness: An examination of internal team environment, shared leadership, and cohesion influences. *Team Performance Management, 19,* 34–56. http://dx.doi.org/10.1108/13527591311312088

DiStefano, C., & Morgan, G. (2014). A comparison of diagonal weighted least squares robust estimation techniques for ordinal data. *Structural Equation Modeling, 21,* 425–438. http://dx.doi.org/10.1080/10705511.2014.915373

Dunkley, D. M., Ma, D., Lee, I. A., Preacher, K. J., & Zuroff, D. C. (2014). Advancing complex explanatory conceptualizations of daily negative and positive affect: Trigger and maintenance coping action patterns. *Journal of Counseling Psychology, 61,* 93–109. http://dx.doi.org/10.1037/a0034673

Eldesouky, L. (2012). Openness to experience and health: A review of the literature. *The Yale Review of Undergraduate Research in Psychology, 5,* 24–42.

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

Fan, X., & Sivo, S. (2005). Sensitivity of fit indices to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling, 12,* 343–367. http://dx.doi.org/10.1207/s15328007sem1203_1

Flack, T., Salmivalli, C., & Idsoe, T. (2011). Peer relations as a source of stress? Assessing affiliation- and status-related stress among adolescents. *European Journal of Developmental Psychology, 8,* 473–489. http://dx.doi.org/10.1080/17405629.2011.558312

Gallagher, M. W., Lopez, S. J., & Preacher, K. J. (2009). The hierarchical structure of well-being. *Journal of Personality, 77,* 1025–1050. http://dx.doi.org/10.1111/j.1467-6494.2009.00573.x

Geiser, C., Keller, B., & Lockhart, G. (2013). First- versus second-order latent growth curve models: Some insights from latent state-trait theory. *Structural Equation Modeling, 20,* 479–503. http://dx.doi.org/10.1080/10705511.2013.797832

Gellert, P., Ziegelmann, J. P., & Schwarzer, R. (2012). Affective and health-related outcome expectancies for physical activity in older adults. *Psychology & Health, 27,* 816–828. http://dx.doi.org/10.1080/08870446.2011.607236

Gershuny, B. S., & Sher, K. J. (1998). The relation between personality and anxiety: Findings from a 3-year prospective study. *Journal of Abnormal Psychology, 107,* 252–262. http://dx.doi.org/10.1037/0021-843X.107.2.252

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

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

Hankonen, N., Konttinen, H., & Absetz, P. (2014). Gender-related personality traits, self-efficacy, and social support: How do they relate to women's waist circumference change? *Journal of Health Psychology, 19,* 1291–1301. http://dx.doi.org/10.1177/1359105313488979

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

Hill, P. L., Payne, B. R., Jackson, J. J., Stine-Morrow, E. A., & Roberts, B. W. (2014). Perceived social support predicts increased conscientiousness during older adulthood. *Journal of Gerontology, Series B, 69,* 543–547. http://dx.doi.org/10.1093/geronb/gbt024

Jackson, L. E., & Gaertner, L. (2010). Mechanisms of moral disengagement and their differential use by right-wing authoritarianism and social dominance orientation in support of war. *Aggressive Behavior, 36,* 238–250. http://dx.doi.org/10.1002/ab.20344

Kaiser, H. F., & Dickman, K. (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika, 27,* 265–267. http://dx.doi.org/10.1007/BF02289635

Karevold, E., Røysamb, E., Ystrom, E., & Mathiesen, K. S. (2009). Predictors and pathways from infancy to symptoms of anxiety and

depression in early adolescence. *Developmental Psychology, 45,* 1051–1060. http://dx.doi.org/10.1037/a0016123

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods & Research, 33,* 188–229. http://dx.doi.org/10.1177/0049124103262065

Kuhn, J., & Holling, H. (2009). Gender, reasoning ability, and scholastic achievement: A multilevel mediation analysis. *Learning and Individual Differences, 19,* 229–233. http://dx.doi.org/10.1016/j.lindif.2008.11.007

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

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.

Lee, S., & Hershberger, S. (1990). A simple rule for generating equivalent models in covariance structure modeling. *Multivariate Behavioral Research, 25,* 313–334. http://dx.doi.org/10.1207/s15327906mbr2503_4

Liao, J., O'Brien, A., Jimmieson, N., & Restubog, S. (2015). Predicting transactive memory system in multidisciplinary teams: The interplay between team and professional identities. *Journal of Business Research, 68,* 965–977.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9,* 151–173. http://dx.doi.org/10.1207/S15328007SEM0902_1

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

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

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. http://dx.doi.org/10.1037/0033-2909.109.3.502

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin, 114,* 185–199. http://dx.doi.org/10.1037/0033-2909.114.1.185

MacCallum, R. C., Widaman, K. F., Preacher, K. J., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research, 36,* 611–637. http://dx.doi.org/10.1207/S15327906MBR3604_06

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4,* 84–99. http://dx.doi.org/10.1037/1082-989X.4.1.84

Mairet, K., Boag, S., & Warburton, W. (2014). How important is temperament? The relationship between coping styles, early maladaptive schemas, and social anxiety. *International Journal of Psychology & Psychological Therapy, 14,* 171–190.

Malmberg, L.-E., & Little, T. D. (2007). Profiles of ability, effort, and difficulty: Relationships with worldviews, motivation, and adjustment. *Learning and Instruction, 17,* 739–754. http://dx.doi.org/10.1016/j.learninstruc.2007.09.014

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

Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology, 10,* 85–110. http://dx.doi.org/10.1146/annurev-clinpsy-032813-153700

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

Martin, M. J., McCarthy, B., Conger, R. D., Gibbons, F. X., Simons, R. L., Cutrona, C. E., & Brody, G. H. (2011). The enduring significance of racism: Discrimination and delinquency among black American youth. *Journal of Research on Adolescence, 21,* 662–676. http://dx.doi.org/10.1111/j.1532-7795.2010.00699.x

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

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.

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

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

Myung, I. J., & Pitt, M. A. (1998). Issues in selecting mathematical models of cognition. In J. Grainger & A. M. Jacobs (Eds.), *Localist connectionist approaches to human cognition* (pp. 327–355). Mahwah, NJ: Erlbaum.

Nasser, F., & Takahashi, T. (2003). The effect of using item parcels on ad hoc goodness-of-fit indexes in confirmatory factor analysis: An example using Sarason's reaction to tests. *Applied Measurement in Education, 16,* 75–97. http://dx.doi.org/10.1207/S15324818AME1601_4

Nasser, F., & Wisenbaker, J. (2003). A Monte Carlo study investigating the impact of item parceling on measures of fit in confirmatory factor analysis. *Educational and Psychological Measurement, 63,* 729–757. http://dx.doi.org/10.1177/0013164403258228

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

Nouwen, A., Urquhart Law, G., Hussain, S., McGovern, S., & Napier, H. (2009). Comparison of the role of self-efficacy and illness representations in relation to dietary self-care and diabetes distress in adolescents with type 1 diabetes. *Psychology & Health, 24,* 1071–1084. http://dx.doi.org/10.1080/08870440802254597

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

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

Owuamalam, C., Issmer, C., Zagefka, H., Klaben, M., & Wagner, U. (2014). Why do members of disadvantaged groups strike back at perceived negativity towards the in-group? *Journal of Community & Applied Social Psychology, 24,* 249–264. http://dx.doi.org/10.1002/casp.2165

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

Preacher, K. J., & Merkle, E. C. (2012). The problem of model selection uncertainty in structural equation modeling. *Psychological Methods, 17,* 1–14. http://dx.doi.org/10.1037/a0026804

Preacher, K. J., Zhang, G., Kim, C., & Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. *Multivariate Behavioral Research, 48,* 28–56. http://dx.doi.org/10.1080/00273171.2012.710386

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

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist, 65,* 1–12. http://dx.doi.org/10.1037/a0018326

Rogers, W. M., & Schmitt, N. (2004). Parameter recovery and model fit using multidimensional composites: A comparison of four empirical parceling algorithms. *Multivariate Behavioral Research, 39,* 379–412. http://dx.doi.org/10.1207/S15327906MBR3903_1

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48,* 1–36. http://dx.doi.org/10.18637/jss.v048.i02

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

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

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464. http://dx.doi.org/10.1214/aos/1176344136

Segrin, C., Woszidlo, A., Givertz, M., & Montgomery, N. (2013). Parent and child traits associated with overparenting. *Journal of Social and Clinical Psychology, 32,* 569–595. http://dx.doi.org/10.1521/jscp.2013.32.6.569

Sexton, K. A., Norton, P. J., Walker, J. R., & Norton, G. R. (2003). Hierarchical model of generalized and specific vulnerabilities in anxiety. *Cognitive Behaviour Therapy, 32,* 82–94. http://dx.doi.org/10.1080/16506070302321

Sierau, S., & Herzberg, P. (2012). Conflict resolution as a dyadic mediator: Considering the partner perspective on conflict resolution. *European Journal of Personality, 26,* 221–232. http://dx.doi.org/10.1002/per.828

Smith, A. (2013). *The relationship of personality to entrepreneurial performance: An examination of openness to experience facets*. Unpublished dissertation. University of Tennessee, Knoxville.

Steiger, J. H. (2002). When constraints interact: A caution about reference variables, identification constraints, and scale dependencies in structural equation modeling. *Psychological Methods, 7,* 210–227. http://dx.doi.org/10.1037/1082-989X.7.2.210

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

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

Sterba, S. K., & Pek, J. (2012). Individual influence on model selection. *Psychological Methods, 17,* 582–599. http://dx.doi.org/10.1037/a0029253

Sterba, S. K., & Rights, J. D. (in press). Accounting for parcel-allocation variability in practice: Combining sources of uncertainty and choosing the number of allocations. *Multivariate Behavioral Research.*

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

Suarez, L., Bennett, S., Goldstein, C., & Barkow, D. (2008). Understanding anxiety disorders from a "triple vulnerability" framework. In M. Antony & M. Stein (Eds.), *Oxford handbook of anxiety and related disorders* (pp. 153–172). New York, NY: Oxford.

Thoits, P. (1984). Explaining distributions of psychological vulnerability: Lack of social support in the face of life stress. *Social Forces, 63,* 453–481. http://dx.doi.org/10.1093/sf/63.2.453

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods, 17,* 228–243. http://dx.doi.org/10.1037/a0027127

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

Williams, P., Rau, H., Cribbet, M., & Gunn, H. (2009). Openness to experience and stress regulation. *Journal of Research in Personality, 43,* 777–784. http://dx.doi.org/10.1016/j.jrp.2009.06.003

Winkler, E., Busch, C., Clasen, J., & Vowinkel, J. (2015). Changes in leadership behaviors predict changes in job satisfaction and well-being in low-skilled work: A longitudinal investigation. *Journal of Leadership & Organizational Studies, 22,* 72–87. http://dx.doi.org/10.1177/1548051814527771

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

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

Zampetakis, L., Kafetsios, K., Bouranta, N., Dewett, T., & Moustakis, V. (2009). On the relationship between emotional intelligence and entrepreneurial attitudes and intentions. *International Journal of Entrepreneurial Behaviour & Research, 15,* 595–618. http://dx.doi.org/10.1108/13552550910995452

Zheng, C., Gaumer Erickson, A., Kingston, N. M., & Noonan, P. M. (2014). The relationship among self-determination, self-concept, and academic achievement for students with learning disabilities. *Journal of Learning Disabilities, 47,* 462–474. http://dx.doi.org/10.1177/0022219412469688

(*Appendices follow*)

## Appendix A

We begin with the item-level common factor measurement model *in the sample* (MacCallum & Tucker, 1991):

$$\mathbf{S}_i = \mathbf{\Lambda}_i\mathbf{C}_{cc_i}\mathbf{\Lambda}_i' + \mathbf{\Lambda}_i\mathbf{C}_{cu_i}\mathbf{\Psi}_i' + \mathbf{\Psi}_i\mathbf{C}_{uc_i}\mathbf{\Lambda}_i' + \mathbf{\Psi}_i\mathbf{C}_{uu_i}\mathbf{\Psi}_i' \quad \text{(A1)}$$

As defined in the article text, $\mathbf{\Lambda}_i$ is a common factor loading matrix, $\mathbf{\Psi}_i$ is a diagonal matrix of unique factor loadings, and $\mathbf{C}_{cc_i}$ is a common factor covariance matrix. Here we define $\mathbf{C}_{cu_i}$ and $\mathbf{C}_{uc_i}$ as containing covariances of common and unique factors and $\mathbf{C}_{uu_i}$ as containing correlations among unique factors. In the population, elements of $\mathbf{C}_{cu_i}$ and $\mathbf{C}_{uc_i}$ would be 0 and $\mathbf{C}_{uu_i}$ would be an identity matrix. But in a sample drawn from this population, nonzero elements of $\mathbf{C}_{cu_i}$ and $\mathbf{C}_{uc_i}$ and nonzero off-diagonal elements of $\mathbf{C}_{uu_i}$ arise purely due to sampling variability and contribute to lack of fit due to sampling error. This lack of fit due to sampling error is represented by $\mathbf{\Delta}_{SE1_i}$ in Equation (1), (as in MacCallum, 2013; MacCallum & Tucker, 1991; MacCallum et al., 1999).

Now we modify this framework in three new ways: (a) we add a structural model; (b) in the structural model, we allow for misfit arising due to sampling error; and (c) we allow for model error specifically from the structural model.

$$\mathbf{S}_i = \mathbf{\Lambda}_i\mathbf{C}_{cc_i}\mathbf{\Lambda}_i' + \mathbf{\Lambda}_i(\mathbf{I}-\mathbf{B}_i)^{-1}\mathbf{C}_{\zeta u_i}\mathbf{\Psi}_i' + \mathbf{\Psi}_i\mathbf{C}_{u\zeta_i}(\mathbf{I}-\mathbf{B}_i')^{-1}\mathbf{\Lambda}_i' + \mathbf{\Psi}_i\mathbf{C}_{uu_i}\mathbf{\Psi}_i' \quad \text{(A2)}$$

$$\mathbf{C}_{cc_i} = (\mathbf{I}-\mathbf{B}_i)^{-1}\mathbf{\Omega}_i(\mathbf{I}-\mathbf{B}_i')^{-1} + (\mathbf{I}-\mathbf{B}_i)^{-1}\mathbf{C}_{offset_i}(\mathbf{I}-\mathbf{B}_i')^{-1} + \mathbf{\Delta}_{ME_i} \quad \text{(A3)}$$

Here—unlike in previous versions of this framework—the covariance matrix among common factors, $\mathbf{C}_{cc_i}$, is structured. As defined in the article text, $\mathbf{B}_i$ is a matrix of regressions among common factors. Here we define $\mathbf{C}_{\zeta u_i}$ and $\mathbf{C}_{u\zeta_i}$ as containing covariances of unique factors and *residuals* of common factors. To relate Equations (A1) and (A2), note that $\mathbf{C}_{cu_i} = (\mathbf{I}-\mathbf{B}_i)^{-1}\mathbf{C}_{\zeta u_i}$ and $\mathbf{C}_{uc_i} = \mathbf{C}_{u\zeta_i}(\mathbf{I}-\mathbf{B}_i')^{-1}$. In Equation (A3), $\mathbf{\Omega}_i$ is a $q \times q$ matrix of residual covariances among common factors and $\mathbf{C}_{offset_i}$ represents discrepancies, due to sampling, in constrained elements of $\mathbf{\Omega}_i$ which can contribute to lack of fit due to sampling error. This lack of fit (due to sampling error) arising from the structural model is represented by $\mathbf{\Delta}_{SE2_i}$ in Equation (2). If sampling error goes to 0, then this term $(\mathbf{I}-\mathbf{B}_i)^{-1}\mathbf{C}_{offset_i}(\mathbf{I}-\mathbf{B}_i')^{-1}$ drops out. Lastly, $\mathbf{\Delta}_{ME_i}$ was defined in the article text.

## Appendix B

One way to represent a likelihood ratio difference test statistic (here denoted $T$) for multivariate normally distributed, mean-deviated manifest variables is as follows. $j$ is the number of manifest parcels. $N$ is sample size. Superscripts $a$ and $b$ refer to Models $a$ or $b$. Superscript $-1$ denotes an inverse. $\mathbf{S}_p$ is the observed sample parcel-level covariance matrix. $\hat{\Sigma}_p^a$ and $\hat{\Sigma}_p^b$ are the model-implied sample parcel-level covariance structures for Models $a$ and $b$.

$$\begin{aligned} T &= (N-1)\left(\left(\ln\left|\hat{\Sigma}_p^a\right| - \ln|\mathbf{S}_p| + tr\left(\hat{\Sigma}_p^{a-1}\mathbf{S}_p\right) - j\right) - \left(\ln\left|\hat{\Sigma}_p^b\right| - \ln|\mathbf{S}_p| + tr\left(\hat{\Sigma}_p^{b-1}\mathbf{S}_p\right) - j\right)\right) \\ T &= (N-1)\left(\left(\ln\left|\hat{\Sigma}_p^a\right| - \ln|\mathbf{S}_p| + tr\left(\hat{\Sigma}_p^{a-1}\mathbf{S}_p\right)\right) - \left(\ln\left|\hat{\Sigma}_p^b\right| - \ln|\mathbf{S}_p| + tr\left(\hat{\Sigma}_p^{b-1}\mathbf{S}_p\right)\right)\right) \\ T &= (N-1)\left(\ln\left|\hat{\Sigma}_p^a\right| - \ln\left|\hat{\Sigma}_p^b\right| + tr\left(\hat{\Sigma}_p^{a-1}\mathbf{S}_p\right) - tr\left(\hat{\Sigma}_p^{b-1}\mathbf{S}_p\right)\right) \end{aligned} \quad \text{(B1)}$$

Now we substitute the Equation (6) expression for $\mathbf{S}_p$ into Equation (B1). $\mathbf{S}_p$ is the same for Models $a$ and $b$. However, the contents of matrices to the right of the equals sign in Equation (6) differ for Models $a$ and $b$, and so are designated as such below. $\tilde{\mathbf{\Sigma}}_p^a$ and $\tilde{\mathbf{\Sigma}}_p^b$ are the model-implied population parcel-level covariance structures for Models $a$ and $b$; other terms were defined in the article.

$$T = (N-1)\left(\ln\left|\hat{\Sigma}_p^a\right| - \ln\left|\hat{\Sigma}_p^b\right| + tr\left(\hat{\Sigma}_p^{a-1}\left(\tilde{\mathbf{\Sigma}}_p^a + \mathbf{\Delta}_{ME_p}^a + \mathbf{\Delta}_{SE_p}^a\right)\right) - tr\left(\hat{\Sigma}_p^{b-1}\left(\tilde{\mathbf{\Sigma}}_p^b + \mathbf{\Delta}_{ME_p}^b + \mathbf{\Delta}_{SE_p}^b\right)\right)\right) \quad \text{(B2)}$$

Equation (B2) is also given in Equation (7). In the Case I situation, Equation (B2) simplifies to

$$\begin{aligned} T &= (N-1)\left(\ln\left|\hat{\Sigma}_p\right| - \ln\left|\hat{\Sigma}_p\right| + tr\left(\hat{\Sigma}_p^{-1}(\tilde{\mathbf{\Sigma}}_p + \mathbf{\Delta}_{ME_p} + \mathbf{\Delta}_{SE_p})\right) - tr\left(\hat{\Sigma}_p^{-1}(\tilde{\mathbf{\Sigma}}_p + \mathbf{\Delta}_{ME_p} + \mathbf{\Delta}_{SE_p})\right)\right) \\ &= 0 \end{aligned} \quad \text{(B3)}$$

In the Case II situation (where $\hat{\Sigma}_p^a \to \tilde{\mathbf{\Sigma}}_p^a$, $\hat{\Sigma}_p^b \to \tilde{\mathbf{\Sigma}}_p^b$, $\mathbf{\Delta}_{SE_p}^a \to 0$, $\mathbf{\Delta}_{SE_p}^b \to 0$), Equation (B2) simplifies to

$$\begin{aligned} T &= (N-1)\left(\ln\left|\hat{\Sigma}_p^a\right| - \ln\left|\hat{\Sigma}_p^b\right| + tr\left(\mathbf{I} + \hat{\Sigma}_p^{a-1}\mathbf{\Delta}_{ME_p}^a\right) - tr\left(\mathbf{I} + \hat{\Sigma}_p^{b-1}\mathbf{\Delta}_{ME_p}^b\right)\right) = \\ T &= (N-1)\left(\ln\left|\hat{\Sigma}_p^a\right| - \ln\left|\hat{\Sigma}_p^b\right| + tr\left(\hat{\Sigma}_p^{a-1}\mathbf{\Delta}_{ME_p}^a\right) - tr\left(\hat{\Sigma}_p^{b-1}\mathbf{\Delta}_{ME_p}^b\right)\right) \end{aligned} \quad \text{(B4)}$$

Received March 10, 2015
Revision received September 25, 2015
Accepted October 3, 2015 ■