Psychological Methods

Population Performance of SEM Parceling Strategies Under Measurement and Structural Model Misspecification

Mijke Rhemtulla

Online First Publication, February 1, 2016. http://dx.doi.org/10.1037/met0000072

Psychological Methods

1082-989X/16/$12.00 http://dx.doi.org/10.1037/met0000072

Population Performance of SEM Parceling Strategies Under Measurement and Structural Model Misspecification

Mijke Rhemtulla
University of Amsterdam

Previous research has suggested that the use of item parcels in structural equation modeling can lead to biased structural coefficient estimates and low power to detect model misspecification. The present article describes the population performance of items, parcels, and scales under a range of model misspecifications, examining structural path coefficient accuracy, power, and population fit indices. Results revealed that, under measurement model misspecification, any parceling scheme typically results in more accurate structural parameters, but less power to detect the misspecification. When the structural model is misspecified, parcels do not affect parameter accuracy, but they do substantially elevate power to detect the misspecification. Under particular, known measurement model misspecifications, a parceling scheme can be chosen to produce the most accurate estimates. The root mean square error of approximation and the standardized root mean square residual are more sensitive to measurement model misspecification in parceled models than the likelihood ratio test statistic.

Keywords: structural equation models, parcels, misspecification, power, fit indices

When fitting a structural equation model with a large number of variables, it can be useful to reduce model complexity by creating parcels. Parcels are formed by summing or averaging scores on two or more indicators of a latent factor, with the goal of reducing the number of indicators of the latent factor. This practice is common but controversial. In particular, proponents of parceling have noted that model fit tends to improve, whereas critics have argued that this improved fit can mask serious model misspecification. In particular, simulation studies have found that misspecification in the measurement model (e.g., correlated residuals and cross-loadings) can lead to the unfortunate combination of biased estimates and good fit. In the present article, I examine the effect of parceling on three types of misspecification, including measurement model misspecification within a latent variable (e.g., unmodeled method variance), measurement model misspecification across latent variables (e.g., unmodeled cross-loadings), and structural model misspecification (e.g., misdirected causal arrows). I examine asymptotic bias in structural parameter estimates, power to detect model misspecification, and population fit statistics in models based on items and parcels.

My goal is to outline under what conditions different strategies—including using the original items, two types of parcels, and scale scores—lead to desirable combinations of bias and power. Studying population performance reveals the clearest possible picture of these effects.

Correspondence concerning this article should be addressed to Mijke Rhemtulla, Department of Psychology, Programme Group Psychological Methods, University of Amsterdam, Weesperplein 4, 1018XA Amsterdam, the Netherlands. E-mail: m.rhemtulla@uva.nl

What Parcels Are Used For

When the research goal is to assess the measurement properties of a scale, parceling is never recommended for the simple reason that it is impossible to study the properties of individual items once they are parceled. When the research goal is to assess relations among constructs, however, it may be appropriate to simplify the measurement model by combining subsets of items into parcels, or by summing all items to create scale scores.

Parceling results in smaller models, indicators with better distributional and psychometric properties, and, frequently, improved model fit. These results are typically seen as benefits of using parcels, and are commonly cited as justifications for modeling parcels rather than items (Bandalos & Finney, 2001; Little, Cunningham, Shahar, & Widaman, 2002; Matsunaga, 2008; Williams & O'Boyle, 2008). But each of these perceived benefits has been disputed.

One of the most common reasons for parceling is to reduce the ratio of variables to sample size (Bandalos & Finney, 2001; Williams & O'Boyle, 2008). Parceling also dramatically reduces the number of model degrees of freedom. One benefit of a smaller model is a more accurate test statistic, because the maximum likelihood chi-square test statistic is upwardly biased when large models are fit to small samples (Anderson & Gerbing, 1984; Boomsma, 1982; Marsh, Hau, Balla, & Grayson, 1998; Moshagen, 2012). Another possible benefit of a smaller model is a greater likelihood that the model will converge to a stable solution (Little et al., 2002; West, Finch, & Curran, 1995). Although simulation studies have shown that convergence problems are actually *less* likely to arise with more indicators per latent factor (Anderson & Gerbing, 1984; Boomsma, 1982; Marsh et al., 1998), these simulations assume that the fitted model is true in the population. It remains possible that parceling increases the likelihood of model convergence by correcting some of the misspecification in an item-level model.

A second common justification for parceling is that it results in variables with better distributional properties (Bandalos, 2002; Bandalos & Finney, 2001; Little et al., 2002; Nasser & Takahashi, 2003; West et al., 1995; Williams & O'Boyle, 2008); that is, parcel data tend to be more continuous and more normal than item data. The default estimation method in all structural equation modeling (SEM) software is normal-theory maximum likelihood, which assumes continuous normal data. To the extent that this assumption is violated, default standard errors and test statistics will be incorrect, and parameters will be underestimated when the data have few categories (Babakus, Ferguson, & Jöreskog, 1987; Green, Akey, Fleming, Hershberger, & Marquis, 1997; Muthén & Kaplan, 1985). Bandalos (2002) found that parceling can ameliorate these estimation problems. On the other hand, all SEM software has implemented robust standard error and test statistic computation for both continuous nonnormal data (i.e., maximum likelihood with robust standard errors and scaled test statistic; Satorra & Bentler, 1994) and for ordinal data (i.e., unweighted or diagonally weighted least squares estimation with mean-and-variance adjusted test statistic; B. O. Muthén, 1993; B. O. Muthén, du Toit, & Spisic, 1997). These methods tend to perform well with ordinal and nonnormal data except under extreme violations of normality (Bandalos, 2008; Rhemtulla, Brosseau-Liard, & Savalei, 2012); thus, this particular justification for parceling may be insufficient (Yang, Nay, & Hoyle, 2010).

A third common justification for parceling is that parcels tend to have better psychometric properties than items (Bandalos & Finney, 2001; Williams & O'Boyle, 2008). Parcels are more reliable, and thus have higher standardized factor loadings and less residual variance (Little, Rhemtulla, Gibson, & Schoemann, 2013; Yuan, Bentler, & Kano, 1997). Many authors have argued that these properties translate to more stable parameter estimates and better convergence for parceled models than item-level models (Landis, Beal, & Tesluk, 2000; Little et al., 2002; Matsunaga, 2008). MacCallum, Widaman, Zhang, & Hong (1999) found that these properties result in higher convergence rates, more accurate parameter estimates, and lower sampling variability. However, the same study also showed that, holding reliability constant, having more indicators per factor leads to similar benefits—so there is evidence that parceling should lead to better performance (because parcels are more reliable) *and* that it should lead to worse performance (because it is better to have more indicators). Marsh et al. (1998) studied this trade-off directly and found no difference in model convergence, parameter accuracy, or efficiency of factor correlations whether parcels or items were used (see also Alhija & Wisenbaker, 2006). Thus, there is no compelling evidence that model convergence, accuracy, or efficiency is actually improved by the better psychometric properties of parcels.

A final frequently cited reason for using parcels is improved model fit (Bagozzi & Heatherton, 1994; Gribbons & Hocevar, 1998; Landis et al., 2000; Takahashi & Nasser, 1996; Thompson & Melancon, 1996). This is the most contentious rationale for parcel use, for two reasons. First, parcels can mask misspecification in measurement model parameters, leading to biased estimates of factor loadings (Hall, Snell, & Foust, 1999; Kim & Hagtvet, 2003) and measurement invariance (Meade & Kroustalis, 2006). But even the strongest proponents of parceling agree that parcels should not be used if the goal is to investigate the measurement properties of a set of items (e.g., Little et al., 2002, 2013). Second, parceling may mask measurement model misspecification, such that bias in structural model parameter estimates goes unnoticed (Bandalos, 2002, 2008; Hall et al., 1999; Rogers & Schmitt, 2004). I consider the evidence for this claim after briefly describing the theoretical framework.

The Effect of Parcels on Model Misspecification

Parcels are created by summing or averaging scores on two or more items to create new variables, to which a model is fitted. The vector of parcel scores for an individual, $\mathbf{X}_{parcel,i} = \mathbf{A}\mathbf{X}_{item,i}$, is a function of the $k \times 1$ vector of item scores and a $p \times k$ allocation matrix, $\mathbf{A}$, which specifies which items belong to which parcels, where k is the number of items and p is the number of parcels. In a structural equation model, item scores are modeled as functions of a set of latent factors: $\mathbf{X}_{items,i} = \mathbf{\Lambda}_{items}\,\boldsymbol{\eta} + \boldsymbol{\varepsilon}_{items,i}$, where $\mathbf{\Lambda}_{items}$ is a $k \times m$ factor loading matrix relating k items to m latent factors, $\boldsymbol{\eta}$ is an $m \times 1$ vector of latent factors, and $\boldsymbol{\varepsilon}_{items,i}$ is a $k \times 1$ vector of item residuals (Jöreskog & Sörbom, 1996).

Let $\mathbf{\Lambda}_{parcels} = \mathbf{A}\mathbf{\Lambda}_{items}$ be the $p \times m$ parcel factor-loading matrix, and $\boldsymbol{\varepsilon}_{parcels,i} = \mathbf{A}\boldsymbol{\varepsilon}_{items,i}$ be the $p \times 1$ vector of parcel residuals, then the parcel level measurement model is $\mathbf{X}_{parcels,i} = \mathbf{\Lambda}_{parcels}\,\boldsymbol{\eta} + \boldsymbol{\varepsilon}_{parcels,i}$. For both item and parcel models, the latent factors are related to each other via a structural model, $\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\zeta}$, where $\mathbf{B}$ is an $m \times m$ matrix of regression coefficients that relate latent variables to each other and $\boldsymbol{\zeta}$ is an $m \times 1$ vector of latent variable residuals. The model implied covariance matrix of items is $\mathbf{\Sigma}_{items}(\boldsymbol{\theta}) = \mathbf{\Lambda}_{items}(\mathbf{I} - \mathbf{B})^{-1}\,\mathbf{\Psi}(\mathbf{I} - \mathbf{B}')^{-1}\mathbf{\Lambda}'_{items} + \mathbf{\Theta}_{items}$, where $\mathbf{\Psi}$ is an $m \times m$ latent variable covariance matrix, $\mathbf{\Theta}_{items}$ is a $k \times k$ residual covariance matrix, and $\mathbf{I}$ is the identity matrix. The corresponding covariance matrix of parcels is $\mathbf{\Sigma}_{parcels}(\boldsymbol{\theta}) = \mathbf{\Lambda}_{parcels}(\mathbf{I} - \mathbf{B})^{-1}\,\mathbf{\Psi}(\mathbf{I} - \mathbf{B}')^{-1}\mathbf{\Lambda}'_{parcels} + \mathbf{\Theta}_{parcels}$, where $\mathbf{\Theta}_{parcels} = \mathbf{A}\mathbf{\Theta}_{items}\mathbf{A}'$ is a $p \times p$ matrix of parcel residuals. Thus, measurement model parameters in $\mathbf{\Lambda}$ and $\mathbf{\Theta}$ differ for items and parcels, whereas structural model parameters in $\mathbf{B}$ and $\mathbf{\Psi}$ do not.

Although $\mathbf{B}$ and $\mathbf{\Psi}$ are not affected by parceling, constraints placed on $\hat{\mathbf{\Lambda}}$ and $\hat{\mathbf{\Theta}}$ for the purposes of estimating the model can be more or less appropriate for parceled versus item models. Differential misfit in the measurement model means that the latent variables themselves are different, which leads to differences between $\hat{\mathbf{B}}_{items}$ and $\hat{\mathbf{B}}_{parcels}$, and between $\hat{\mathbf{\Psi}}_{items}$ and $\hat{\mathbf{\Psi}}_{parcels}$. For example, if several items that share residual covariance are allocated to the same parcel, the shared variance that belonged to off-diagonal elements of $\mathbf{\Theta}_{items}$ will belong to the diagonal of $\mathbf{\Theta}_{parcels}$. If this residual covariance is not modeled in $\mathbf{\Theta}_{items}$, then some of that shared variance will become part of the latent variable in the item model but not the parceled model, and the item model will show worse fit than the parceled model. On the other hand, if several indicators of a latent factor all share a small amount of variance with another latent factor, and these items are allocated to the same parcel, the resulting parcel will have an even greater tendency to cross-load. That is, several small misspecifications in $\hat{\mathbf{\Lambda}}_{items}$ will be repackaged into one big misspecification in $\hat{\mathbf{\Lambda}}_{parcels}$. By adjusting or reallocating measurement misfit, parceling can change the latent variables that are identified, and thereby affect structural model parameter estimates.

Research investigating the effect of parceling on bias and model fit have compared two kinds of parcels. *Isolated parcels* combine items that share the same sources of variance; for example, items

that all load on an (unmodeled) group factor are placed in the same parcel. By combining similar items together, a set of isolated parcels are maximally different from each other. In contrast, *distributed parcels* distribute items that share the same sources of variance across parcels, creating a set of parcels that are maximally similar to each other.[1] A common factor modeled on a set of isolated parcels will have smaller variance than one modeled on a set of distributed parcels, because isolated parcels have less in common with each other than distributed parcels. These strategies will have different effects on structural parameter estimates and the degree of model misfit; which strategy is preferable depends on the type of measurement model misspecification.

Simulation studies have confirmed that when a subset of item-level indicators of a factor share a secondary source of variance (e.g., method variance) that is unrelated to other variables in the model, isolated parceling can isolate that secondary variance and remove it from the structural model (Hall et al., 1999). In this situation, both distributed parcels and items should result in more bias than isolated parcels, though no research has yet examined how items compare to parcels in this context. In terms of model fit, Hall et al. (1993) found that despite differences in bias, both isolated and distributed parcels resulted in well-fitting models: Isolated parcels fit well because the secondary variance is relegated to a parcel residual, and distributed parcels fit well because the secondary variance is shared across all parcels and thus becomes part of the latent variable. In contrast, a model based on the original items would be expected to fit poorly, because the secondary source of variance is neither isolated in a single indicator, nor is it shared across all of them.

When the secondary variance source is related to other variables in the model (e.g., method variance affects two latent variables), some research has found that isolated parcels can intensify the effect of secondary variance, resulting in worse bias than distributed parcels (Bandalos, 2002, 2008), but other research has found that both isolated and distributed parcels result in similar levels of bias (Hall et al., 1999). Bandalos (2002) found that item-based models result in the same high degree of bias but greater model misfit than isolated parcels. Several studies have found that distributed parcels tend to result in relatively well-fitting models, making it difficult to detect measurement model misspecification (Bandalos, 2002, 2008; Hall et al., 1999; Rogers & Schmitt, 2004).

Although the simulation literature reveals some intuitive findings, it is nonetheless difficult to draw broad conclusions from these results. These studies have typically examined only one or two models with a fixed set of parameter values, making it difficult to generalize results to other types of misspecifications. Moreover, they have frequently simulated ordinal data, making it difficult to distinguish bias due to violations of the normality distribution from that due to parceling (Bandalos, 2008). Other studies have considered empirical parceling schemes, which only partially achieve the effects of isolated or distributed parceling (e.g., Rogers & Schmitt, 2004). Only one study has compared bias resulting from item models with parcel models (Bandalos, 2002). No study has systematically compared item solutions with parceled solutions across multiple types of misspecification, nor has any study examined the effects of parceling on structural model misspecification.

Overview

The present study examines the repercussions of fitting a misspecified latent mediation model to population item, parcel, and scale data, in terms of bias, model fit, and power to detect a range of misspecifications. By exploring the effects of parceling under a range of common model misspecifications, my aim is to reveal general principles of how and why parceling can exacerbate or diminish the effects of model misspecification compared with items and compared with alternate parceling schemes.

Studying population performance has several advantages over Monte Carlo simulation (Reise, Scheines, Widaman, & Haviland, 2013; Rhemtulla, Savalei, & Little, 2014; Savalei, 2012). For one, results are more precise than those based on even thousands of simulated replications.[2] Second, the population noncentrality parameter can be used to compute power for an arbitrary range of sample sizes, giving a precise estimate of the sample size required to achieve sufficient power. Third, because the population approach is less computationally demanding (as no sample data are generated or analyzed), it is possible to investigate a more comprehensive set of parameter values. The reader should keep in mind, however, that the results of population explorations do not include sampling variability; that is, the present results indicate the levels of bias and model fit that one would expect to see on average, but not necessarily in any given finite sample.

In three studies, I consider three ways in which to misspecify a structural equation model. First, the measurement model of any single construct could be wrong if the shared variance among the set of indicators cannot be fully explained by a single latent variable; that is, more than one source of shared variance affects the set of indicators. In addition, none of these sources also affect other variables (latent or observed) in the model. For example, the true model might be a two- (or more) factor model, a bifactor model, a higher order model, or a model with correlated residuals. Given this type of misspecification, isolated parceling combines items that share secondary variance, so that the secondary variance (unshared with anything else in the model) gets relegated to parcel residuals and is thus removed from the structural model. Distrib-

[1] Virtually every parceling strategy that has been proposed is a variant of one of these strategies: Radial parceling (Cattell, 1956; Cattell & Burdsal, 1975), correlational parceling (Landis et al., 2000; Rogers & Schmitt, 2004), item-to-construct balanced parceling (Little et al., 2002), domain-representative parceling (Coffman & MacCallum, 2005; Kishton & Widaman, 1994), and single factor analysis parceling (Mathieu & Farr, 1991) all aim to create maximally similar parcels (i.e., distributed parcels), whereas unidimensional parceling (Kishton & Widaman, 1994), exploratory factor analysis parceling (Landis et al., 2000), and empirically equivalent parceling (Landis et al., 2000) all aim to maximize within-parcel similarity (i.e., isolated parcels).

[2] Population results from a single level of misspecification and sample size in Model 1A were successfully replicated via simulation (1,000 replications with $N = 400$). The simulation yielded parameter estimates that were identical to three decimal places, average sample root mean square error of approximation (RMSEA) and comparative fit index (CFI) that followed the same pattern of results and typically matched to two decimal places, and average standardized root mean square residual (SRMR) values that were higher (due to sampling variability) but also followed the same pattern of results. Power matched to within 1% for scale and parcel models, but item-level power was higher in simulated data due to the known upward bias in the test statistic when the number of manifest variables is large (Moshagen, 2012).

uted parceling, in contrast, distributes items that share secondary variance across multiple parcels, so that the secondary variance becomes part of the common factor variance.

Study 1 considers a two-factor model, in which both sources of variance are equally relevant to the latent construct under investigation, and a one-factor model in which a subset of items are affected by a method factor, which is irrelevant to the construct. In the first case, it is desirable that a modeled factor contains both sources of variance, so distributed parcels are expected to minimize the bias to structural model parameters; the degree of bias produced by isolated parcels relative to items is unclear. In the second case, it is desirable that the method variance not be included in the modeled factor, so isolated parcels are expected to minimize bias, and the degree of bias produced by distributed parcels relative to items is unclear. In both cases, items are expected to result in higher power than parcels (Bandalos, 2002).

Study 2 considers measurement model misspecification that is due to a secondary source of variance that is shared with indicators of another construct. For example, the true model could include cross-loadings, in which indicators of one factor are directly affected by another factor, correlated residuals across indicators of two factors, or a method factor that affects indicators of multiple constructs. When multiple items are affected, isolated parcels combine several small misspecifications into one bigger one. This strategy has been found to result in higher power than distributed parceling (Bandalos, 2002, 2008; Hall et al., 1999), in which small misspecifications are spread across several parcels, diffusing the misspecification. It is not clear based on previous research or theory how the two parceling strategies should compare with each other or to items with respect to bias.

Finally, Study 3 considers misspecification in the structural model, when the measurement model is correctly specified. As explained earlier, misfit stemming from the structural model should not be affected by parceling. Any misspecification in the structural model should remain when parcels are employed; however, power to detect the misspecification may differ (Ledgerwood & Shrout, 2011).

Method

Population and Fitted Models

In each condition, the population generating model introduced some misspecification to the latent mediation model depicted in Figure 1 (top; these misspecifications are described in detail in following sections). The latent mediation model is a simple structural equation model that captures a very common hypothesis in psychological research (e.g., Rucker, Preacher, Tormala, & Petty, 2011). The model is just complex enough to support a range of misspecification types, including correlated residuals, cross-loadings, and structural path misspecification. Standardized primary factor loadings were always .5, reflecting a situation in which researchers may be inclined to form parcels to increase the reliability of the indicators. All other parameters values for the item-level population model were chosen to result in approximately the same degree of misfit in the item-level model across conditions (population root mean square error of approximation [RMSEA] $\approx$ .02[3]). The required sample size to reach 80% power to detect the misspecification in the item-level model by the chi-square test of exact fit, in all conditions, was about 400. In addition, for each model, a single parameter involved in the misspecification was varied continuously over a large range of values to reveal the effect of the degree of misspecification on parameter bias.

Parcel and scale covariance matrices were obtained by pre- and postmultiplying the item covariance matrix by one of three allocation matrices to create three 4-item parcels per factor: *isolated parcels* allocated the first four, next four, and last four indicators of each latent construct to a parcel; *distributed parcels* allocated indicators 1/4/7/10, 2/5/8/11, and 3/6/9/12 of each latent construct to a parcel, and *scales* combined all indicators of each latent factor into a single composite. Proponents of parceling have recommended using exactly three parcels per latent variable because it results in greatest model reduction while still allowing each latent variable to be locally identified (Little, 2013). However, to test the generality of results, all analyses were also conducted with four 3-item parcels per factor. Four isolated parcels were formed by allocating indicators 1–3, 4–6, 7–9, 11–12 of each factor to parcels; four distributed parcels were formed by allocating indicators 1/5/9, 2/6/10, 3/7/11, and 4/8/12 of each factor to parcels. In Studies 1 and 2, results are plotted and discussed in terms of the three-parcel solutions, and any notable differences are discussed in the text. In Study 3, both three- and four-parcel results are plotted.

The test models shown in Figure 1 were fit to each item-level, parcel-level, and scale-level covariance matrix, using the R package *lavaan* (Rosseel, 2012). Note that even when no misspecification is introduced, the scale model is still misspecified: Because the scale model assumes that X, M, and Y are measured without error (i.e., there is no measurement error in the model), both the $X \rightarrow M$ path and the $M \rightarrow Y$ path will be attenuated. As a further result of this attenuation, the indirect $X \rightarrow M \rightarrow Y$ path is unable to account for the covariance between X and Y, so the model does not fit. (For a fuller explanation of the consequences of modeling latent variables using scales, see Cole & Preacher, 2014). Thus, bias, fit, and power in the scale models will reflect a combination of this baseline misspecification plus whatever is due to the additional imposed misspecification. In contrast, the item- and parcel-level latent variable models shown in Figure 1 are correctly specified when no further misspecification is imposed.

Outcomes

Bias. Structural parameter accuracy was assessed by examining the value of the indirect effect of X on Y; that is, the product of the $X \rightarrow M$ coefficient and the $M \rightarrow Y$ coefficient. In keeping with mediation model terminology, I call this indirect path *ab*. The population *ab* was always .16 for every condition except Model 3A, where the $M \rightarrow Y$ path was reversed, resulting in a population indirect effect of 0. Deviation from .16 by .01 in either direction corresponds to 6.25% relative bias, $RB = (ab - .16)/.16$. Relative

[3] Misspecification resulting in a RMSEA of .02 may seem inconsequential, but it is nearly impossible to stipulate local misspecifications (e.g., a pair of cross-loadings) that lead to bigger misspecifications without imposing implausibly high factor loadings or structural coefficients (see Savalei, 2012). Results of the present set of studies reveal that even this small overall degree of misspecification typically leads to consequential bias. Of course, when more extreme misspecification is present, bias and power may be expected to be higher across the board; however, the relative performance of the item, parcel, and scale models is unlikely to be affected.

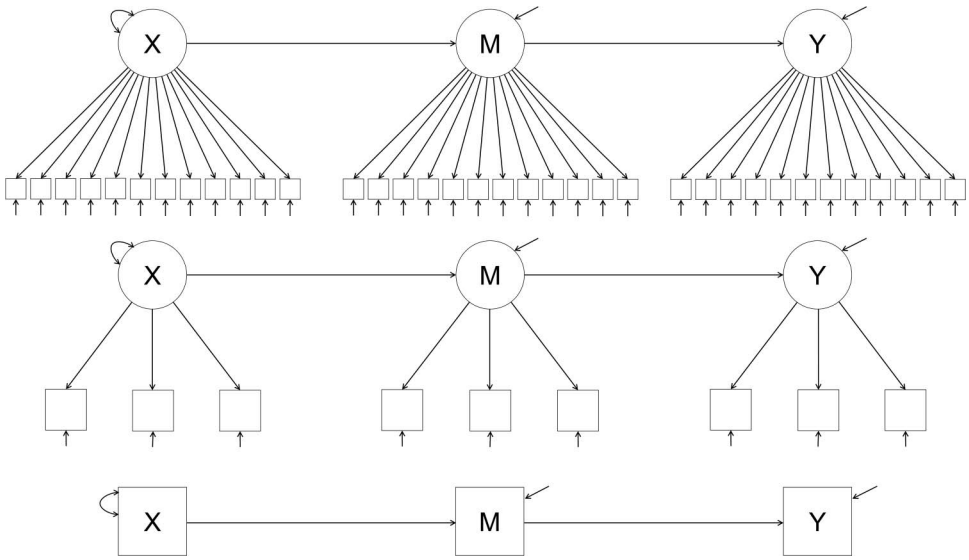

Figure 1. Test models for item data (top), parcel data (middle), and scale data (bottom).

bias cannot be computed for Model 3A as that would require dividing by 0. Raw *ab* values are presented in figures, and relative bias is discussed in-text.

Power. Power to detect model misspecification using the chi-square test of exact fit is a function of the degree of model misspecification (i.e., the minimum of the maximum likelihood fit function, $\hat{F}_{ML}$) and the sample size. Thus, $\hat{F}_{ML}$ obtained from a misspecified model fit to population data can be used to obtain power for any arbitrary sample size, N (Saris & Satorra, 1993; Satorra & Saris, 1985). The noncentrality parameter of the noncentral chi-square distribution, λ, is given by $N \times \hat{F}_{ML}$ (Satorra & Saris, 1985). Power corresponds to the proportion of the noncentral chi-square distribution that falls beyond the critical value for the null hypothesis that the model is correctly specified, $P\{\chi^2(df, \lambda) > c_{.05}\}$, where $c_{.05}$ is the critical value for $\alpha = .05$, and df is the model degrees of freedom.

Fit indices. Three population fit indices are presented as a continuous function of each model misspecification (e.g., as a function of a cross-loading strength). First, the RMSEA captures the degree of discrepancy between the population covariance matrix and the model-reproduced covariance matrix, per degrees of freedom (Browne & Cudeck, 1993). The population RMSEA is given by $RMSEA = \sqrt{\hat{F}_{ML}/df}$ (Steiger, 1990; Steiger & Lind, 1980), of which the sample RMSEA is an unbiased and consistent estimator (Browne & Cudeck, 1993; McDonald, 1989).

Power of the chi-square test of exact fit is equivalent to a test of the null hypothesis that population RMSEA = 0. Power to detect other population RMSEA values (e.g., to test the close-fit hypothesis; MacCallum, Browne, & Sugawara, 1996) is not presented, but can be extrapolated from the patterns shown in the power curves for each model: Each power curve becomes progressively flatter as the null hypothesis value of RMSEA increases. When the criterion RMSEA value is larger than the population RMSEA for the fitted model, the curve for that model converges upon 0, rather than 1.

Second, the population standardized root mean square residual (SRMR) is presented for each fitted model. SRMR is the average standardized residual:

$$SRMR = \sqrt{2\sum_{i=1}^{p}\sum_{j=1}^{i}\left\{(s_{ij} - \hat{\sigma}_{ij})/(s_{ii}s_{jj})^{\frac{1}{2}}\right\}^2/p(p+1)},$$

where s is the sample covariance matrix (in the present study, the population covariance matrix), $\hat{\sigma}$ is the fitted (model-implied) covariance matrix, and p is the number of observed variables (Bentler, 2006).

Finally, Bentler (1990) described a population comparative fit coefficient, $\Delta = 1 - \lambda_{fitted}/\lambda_{null}$, where λ_{fitted} is the noncentrality parameter for the fitted model, and λ_{null} is the noncentrality parameter for the null model. In the null model, the variances of all variables are freely estimated, and all covariances are constrained to 0. The comparative fit index (CFI; Bentler, 1990), normed fit index (NFI; Bentler & Bonett, 1980), and incremental fit index (IFI; Bollen, 1989), are all consistent estimators of this population coefficient. I refer to this fit index as "population CFI," but it can equally be thought of as population NFI or IFI.

Study 1: Misspecification in the Single-Factor Measurement Model

Figure 2 shows the two population models considered in Study 1. In Model 1A, the latent construct X is represented by two correlated factors, one with eight indicators (*X1*) and one with four (*X2*). That is, what the researcher has conceptualized as a unidimensional construct (e.g., *aggression*) actually encompasses two correlated factors (e.g., *proactive* and *reactive aggression*; Poulin & Boivin, 2000). The "true" effect of $X \rightarrow M$ is conceptualized as the total effect of *X1* and *X2* on *M*. Isolated parcels were formed by combining the four indicators of *X2* into a single parcel, and the eight indicators of *X1* into two parcels. Distributed parcels were formed by distributing the four indicators of *X2* across three

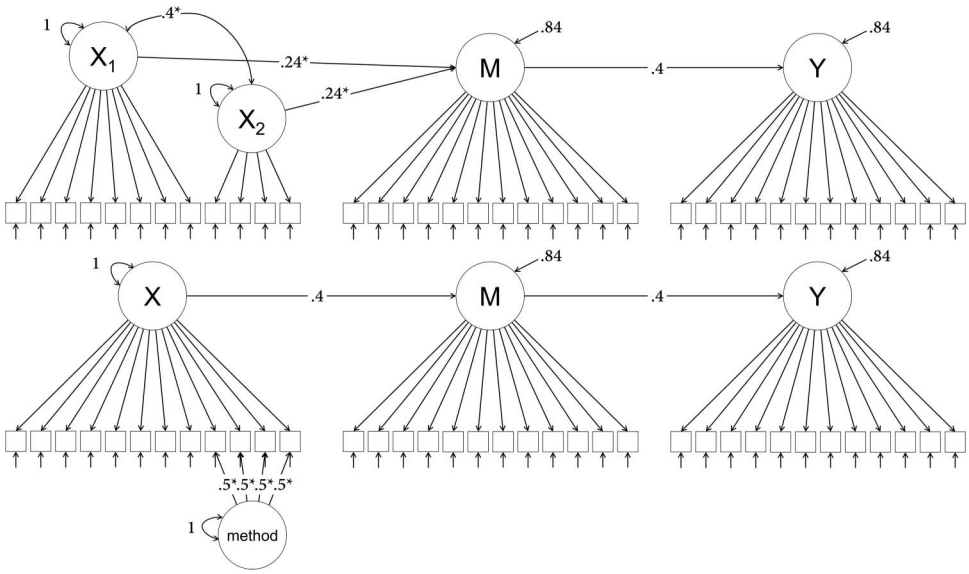

Figure 2. Population Model 1A (upper) and 1B (lower). Isolated parcels combine the four indicators of *X2* (Model 1A) or the four indicators that load on the method factor (Model 1B) into a single parcel. Distributed parcels allocate these four items to three different parcels. All factor loadings are .5. All manifest and latent variables have a total variance of 1. Asterisks denote population values that are varied continuously.

parcels, in combination with the eight indicators of *X1*. In this misspecified model, what is modeled as a single common factor (*X*) is actually composed of two correlated sources of variance (*X1* and *X2*). By allocating all indicators of *X2* to a single parcel, the isolated parceling strategy will relegate unique *X2* variance to that parcel's residual, producing a common factor that represents a mix of *X1* variance and the variance shared among *X1* and *X2*. Distributed parceling, in contrast, will produce a common factor that contains unique variance from both *X1* and *X2*.

Model parameters were chosen such that the proportion of variance in the latent mediator, *M*, accounted for by *X1* and *X2* together was .16. Within this model, two continuously varying effects were examined: First, the correlation between *X1* and *X2* was varied from 0 to 1 in increments of .01. In this manipulation, the regression coefficients of $X1 \rightarrow M$ and $X2 \rightarrow M$ were equal to each other, and their value varied simultaneously with the (*X1*, *X2*) correlation to keep the total $X \rightarrow M$ effect constant. Second, the proportion of the $X \rightarrow M$ effect due to *X1* versus *X2* was varied from 0 (i.e., the entire effect is due to *X2*) to 1 (i.e., the entire effect is due to *X1*). In this second manipulation, the (*X1*, *X2*) correlation was held constant at .4, and the total $X \rightarrow M$ effect was again held constant.

In Model 1B, a secondary method factor accounts for additional shared variance among four of the 12 indicators of Factor *X*. Isolated parcels are formed by combining these four indicators with method variance into a single parcel; distributed parcels distributed these four indicators across three parcels. Because the method factor is unrelated to other variables in the model, isolated parceling is expected to eliminate the misspecification by removing method variance from the structural model. Distributed parceling, in contrast, should bring the method variance into the common factor, and thus introduce bias into the structural model (Hall et al., 1999).

Results: Model 1A

Figure 3 (top left) shows the model-estimated *ab* for the models based on items, isolated and distributed parcels, and scales as a function of the correlation between *X1* and *X2*. The solid gray line at $ab = .16$ indicates the population parameter value when the correct population model is fit. When *X1* and *X2* are perfectly correlated, there is no misspecification (i.e., it is correct to model them as a single factor); the misspecification becomes more severe as the correlation decreases. At almost all values of this correlation, bias in *ab* is most severe in scale data, similar for items and isolated parcels, and benign in distributed parcels. For example, when $\text{cor}(X1, X2) = .4$, the *ab* estimates were .097 for scales (−40% bias), .142 for isolated parcels (−11% bias), .149 for items (−7% bias), and .157 for distributed parcels (−2% bias).

Figure 3 (top right) shows power curves for detecting the model misspecification for each parceling strategy when $\text{cor}(X1, X2) = .4$ as a function of *N*. Items have the highest power to detect the misspecification, attaining 80% power at $N = 400$. Isolated parcels have considerably less power, attaining 80% power at $N = 1{,}081$. Scales have substantially less power, reaching 80% power only at $N = 10{,}976$. Distributed parcels have essentially no power, reaching only 27% power by $N = 10{,}000$.

To understand how bias and misfit arises, consider the source of misspecification for each model. For items, there are two subsets of indicators that are more strongly related to each other than to the others, so a single factor cannot adequately account for their covariances. Because the measurement model is incorrect, the model cannot account for relations between *X* indicators and *M* indicators, leading to model misfit and underestimation of the $X \rightarrow M$ coefficient.

With isolated parcels, one parcel represents *X2* and two parcels represent *X1*. Because all variance unique to *X2* is contained within

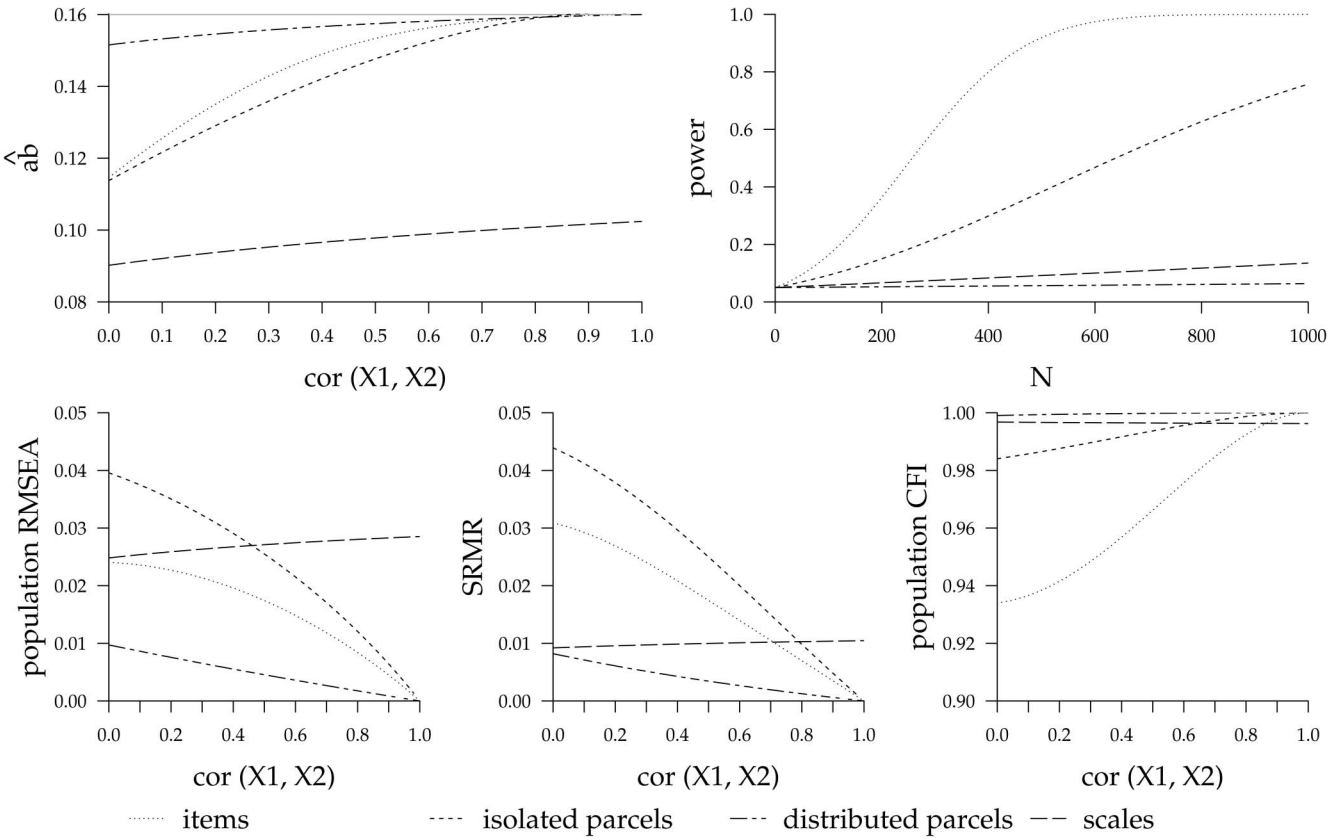

Figure 3. Model 1A results when *X1* → *M* and *X2* →*M* effects are equal. Top left: Estimated values of *ab* for the latent mediator (*M*) as a function of the population covariance between the latent generating factors *X1* and *X2* (*x*-axis) and parceling strategy (separate lines). The solid gray line at *ab* = .16 indicates the population value. Top right: Power to detect model misspecification by sample size (*x*-axis) and parceling strategy (separate lines) when cor(*X1*, *X2*) = .4. Bottom: Population fit indices as a function of the population correlation between the latent generating factors *X1* and *X2* (*x*-axis) and parceling strategy (separate lines).

a single parcel, *X2* variance is relegated to residual variance of that parcel. The unique relation between *X2* and *M* is not captured in the model, leading to both bias (*X* → *M* is underestimated) and misfit (covariance between the parcel representing *X2* and indicators of *M* and *Y* are not accounted for by the model).

With distributed parcels, all three parceled indicators of *X* represent both *X1* and *X2*. As such, the *X* construct represents all shared variance. The small amount of model misfit is caused by a slight imbalance in the parcels—one parcel contained two *X2* items and the others contained just one. When four distributed parcels are created, each containing one *X2* item, the resulting model fit is perfect. However, even though four distributed parcels results in perfect fit, the slight bias in *ab* observed with three distributed parcels remains (the degree of bias is almost identical with three or four distributed parcels). This bias occurs because *X2* variance is underrepresented in each parcel compared with *X1* variance: A parcel containing two *X1* indicators and one *X1* indicator contains a ratio of 4:1 *X1*:*X2* variance.

Finally, as mentioned earlier, bias in the scale model has two additive sources: attenuation of the *ab* path due to measurement error, plus bias due to the additional measurement model misspecification. Bias due to attenuation can be seen at the far right of the top left plot of Figure 3: When cor(*X1*, *X2*) = 1, every latent variable method produces accurate estimates, but the manifest variable approach based on scales results in an underestimate (*ab* = .102; −36% bias). As cor(*X1*, *X2*) decreases, *ab* decreases further. Misfit in the scale model arises because the underestimated *ab* does not fully reproduce the covariance between the *X* and *Y* scales, and the more extreme the bias, the worse the misfit.

The bottom row of Figure 3 shows the population fit indices as a function of cor(*X1*, *X2*). SRMR and RMSEA indicate worse fit for isolated parcels than items. Thus, SRMR and RMSEA appear to be more sensitive to model misspecification in parceled models, even when these parcels result in less bias than items. In contrast, the population CFI is most indicative of item misfit. Distributed parcels produce close to perfect fit by every metric.

Figure 4 (left) shows the model-estimated *ab* as a function of the proportion of the *X* → *M* effect due to *X1*. The solid gray line at *ab* = .16 indicates the population parameter value when the correct population model is fit. Bias in *ab* is always most severe in scale data. Items, isolated parcels, and distributed parcels all display a pattern of bias that decreases as more of the effect is due to *X1*, up to a point of maximum accuracy that differs across the three ways of modeling the latent variable. Isolated parcels result in the most

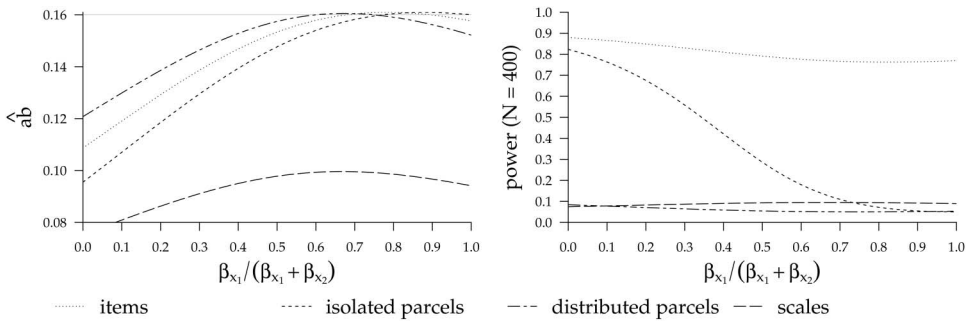

Figure 4. Model 1A results when *X1* → *M* and *X2* → *M* effects are unequal. Left: Estimated values of *ab* for the latent mediator (*M*) as a function of the proportion of the *X* → *M* effect due to *X1* (*x*-axis) and parceling strategy (separate lines). The solid gray line at *ab* = .16 indicates the population value. Right: Power to detect model misspecification as a function of the proportion of the *X* → *M* effect due to *X1* (*x*-axis) and parceling strategy (separate lines). The correlation between *X1* and *X2* is held constant at .4, and the proportion of the variance in *M* due to both *X1* and *X2* is .16.

bias, especially when most of the effect is due to *X2*, distributed parcels result in the least bias until about 65% of the effect is due to *X1*, and the item model estimates are in between. Though the bias of the three models is not very different, power to detect the misspecification differs dramatically across models. The right panel of Figure 4 displays power when $N = 400$ as a function of the proportion of the *X* → *M* effect due to *X1*. When items are modeled, power is consistently high, despite decreasing bias. When isolated parcels are modeled, power decreases in line with the decreasing bias, until, when the entire effect is due to *X1*, there is zero bias in *ab* and the model has perfect fit. When distributed parcels are modeled, there is virtually no power at any point to detect the misspecification.

Further analytical explorations of these effects revealed that bias can be fully accounted for by three interacting factors: (a) cor(*X1*, *X2*)—all methods become increasingly accurate as this correlation increases, as seen in Figure 3; (b) the proportion of the variance in *M* due to *X1* relative to that due to *X2*—when these effects are balanced (as in Figure 3, and in Figure 4 when the *x*-axis value is .5), distributed parcels result in little to no bias, and items and isolated parcels result in slightly more bias; and (c) the number of indicators of *X1* and *X2*—when there are an equal number of indicators, the bias curves in Figure 4 are symmetrical around .5; as there are more indicators of *X1*, these peaks shift to the right (as seen in Figure 4). Given that these factors are not likely to be known, it is difficult to recommend one best method. However, it is clear from all conditions that (a) distributed parcels *tend* to result in the least amount of bias but no power; (b) isolated parcels have power that reflects the degree of bias in the structural parameter, such that biased structural parameter estimates are more likely to be detected; and (c) items have high power regardless of the degree of structural parameter bias, such that even unbiased estimates are likely to result in poor model fit.

These results are more promising than those reported by Bandalos (2008), who studied a very similar model but measured bias according to a different conception of the "true" effect. Here, I consider the true *X* → *M* parameter to be the combined effect of *X1* and *X2* on *M* (e.g., the total effect of proactive and reactive aggression). This approach is appropriate if the researcher is interested in assessing the combined effect of all facets of a complex construct, as is likely given that he is modeling the construct as a single factor. Bandalos considered the true *X* → *M* parameter value to be the single effect of *X1* on *M* (e.g., the effect of proactive aggression alone). Her approach is appropriate under the assumption that the researcher is interested in assessing only one of the facets of a complex construct. A third approach would be to consider the true *X* → *M* parameter to be the effect of a common factor representing variance shared by *X1* and *X2* (e.g., the effect of a higher order aggression factor on which proactive and reactive aggression load). This approach would be appropriate under the assumption that the researcher is interested in assessing the effect of only the variance that is common to all facets of a complex construct. This approach results in more extreme bias for all methods, with very little difference between them, and similar power to that presented in Figure 3.

Results: Model 1B

Figure 5 (top left) shows the model-estimated *ab* for the models based on items, isolated and distributed parcels, and scales as a function of (standardized) loadings on the secondary "method" factor in Model 1B. When method loadings are 0, there is no method variance and therefore no misspecification in any of the latent variable models. As method loadings increase, the misspecification becomes more severe. Bias in the scale model is barely affected by the strength of method loadings; all other models show quite a small degree of bias until the method factor loadings reach about .5 (i.e., the same strength as their loadings on the *X* factor). Of the latent variable models, when method factor loadings are .5, items produce the most bias ($ab = .143$; −10% bias), distributed parcels produce small bias ($ab = .152$; –5% bias), and isolated parcels produce no bias (consistent with Hall et al., 1999).

The top right panel of Figure 5 shows power curves for each parceling strategy when the standardized method factor loadings are .5. The item-level model is the only one with any discernable power to detect the misspecification, reaching 80% power at $N = 366$. Scales require $N = 10{,}361$ to attain 80% power, distributed parcels have only 15% power by $N = 10{,}000$, and isolated parcels have perfect fit, so "power" in this case equals the Type I error rate.

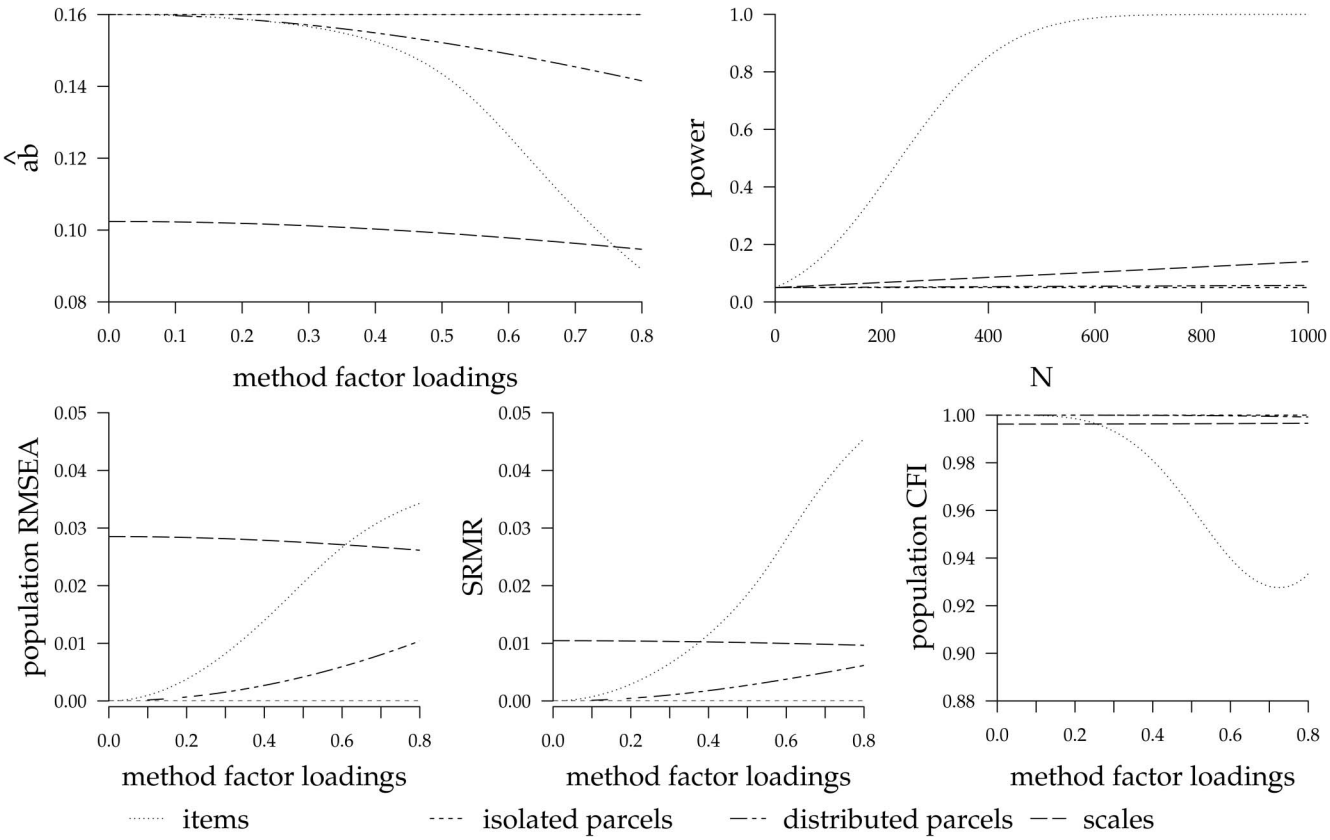

Figure 5. Model 1B results. Top left: Estimated values of *ab* as a function of the strength of standardized factor loadings on the secondary method factor (*x*-axis) and parceling strategy (separate lines). The solid gray line at $ab = .16$ is the population value. Top right: Power to detect model misspecification by sample size (*x*-axis) and parceling strategy (separate lines) when method factor loadings are .5. Bottom: Population fit indices as a function of the strength of standardized factor loadings on the secondary method factor (*x*-axis) and parceling strategy (separate lines).

The bottom row of Figure 5 shows fit indices as a function of the method factor loadings. Of the latent variable models, the item model displays the worst fit, whereas distributed parcels show much better fit and isolated parcels have perfect fit. Misfit in the scale model actually decreases slightly with the increasing misspecification, which counteracts the baseline misspecification of the scale model.

The key to understanding these results is to recall that the method factor is uncorrelated with *M* or *Y*. The isolated parceling strategy is optimal because it allocates all shared method variance to a single residual. The distributed parceling strategy, in contrast, allows method variance to be shared across all indicators of *X*, such that *X* represents both the intended construct and a particular method. This strategy creates no misspecification within the *X* model, because all distributed parcels share both sources of variance. Although the model based on distributed parcels fits well, the relations between *X* and other variables in the structural model are attenuated as a result of the extra method variance in *X*. The present results differ from previous studies in which the method factor affected multiple constructs (e.g., Bandalos, 2002, 2008; Rogers & Schmitt, 2004): When the method factor is not unique to one factor, isolated parceling will not prevent method variance from entering the structural model.[4] Study 2 considers more closely the situation in which secondary variance is related to other variables in the model.

It is worth noting the role of variability in factor loadings and method factor loadings in this model. In this study, all factor loadings and method factor loadings were held constant across items. In the literature on common method variance, a distinction is made between a set of items that are equally affected by method variance ("noncongeneric" items) and those that are unequally affected ("congeneric" items). This literature has found that the effectiveness of techniques for dealing with common method variance depends on this difference (e.g., Richardson, Simmering,

[4] In a variant on Model 1B (not presented), method variance affected a subset of indicators of both *X* and *M*, resulting in a spuriously high $X \rightarrow M$ path due to correlated method variance across the two factors. In this situation, items, distributed parcels, and isolated parcels all resulted in almost identical levels of upward bias in the structural coefficient, consistent with previous research. Isolated parcels resulted in highest power, followed closely by items, whereas distributed parcels had very low power and scales had almost none. These findings are very similar to those resulting from cross-loadings (Study 2)—another situation in which a secondary source of variance is related to other variables in the model.

& Sturman, 2009; Schaller, Patil, & Malhotra, 2015; Simmering, Fuller, Richardson, Ocal, & Atinc, 2015). To investigate whether unequal factor loadings or method factor loadings may affect the differential performance of items and parcels, Models 1A and 1B were replicated with varying factor loadings as well as varying method loadings. The results were virtually indistinguishable from those presented here: Neither absolute levels of bias and power nor the relative performance of each method were affected when factor loadings varied or when common method variance unequally affected items.

Study 2: Misspecification in Multifactor Measurement Model

Study 1 considered two model misspecifications that affect the measurement model of a single factor. But measurement model misspecifications can be more subtle: Even perfectly fitting single-factor measurement models may be misspecified when they are placed in a model with other constructs. Scales are typically validated using single-factor confirmatory factor analysis (CFA) models, and parceling strategies tend to be based on inter-item properties within a scale. But indicators of a latent factor may be related to another construct for reasons unrelated to the primary construct they indicate. For example, the Masculinity scale of the Bem Sex Role Inventory contains an item that queries self-sufficiency (Bem, 1974). A researcher studying the relation between masculinity and work performance would likely find that self-sufficiency is related to work performance over and above its association with masculinity. When such cross-associations are anticipated, what is the best way to deal with them?

I consider three population models. In Model 2A, one indicator of Factor *X* has a positive cross-loading on Factor *M*, and one indicator of Factor *M* has a positive cross-loading on Factor *Y*. In population Model 2B, two indicators of Factors *X* and *M* have positive cross-loadings on Factors *M* and *Y*, respectively (e.g., the masculinity items "self-sufficiency" and "acts as a leader" may both load positively on work performance over and above their relation to masculinity). Population Model 2C is identical to 2B, except one of the two cross-loadings on each factor is negative (e.g., the masculinity item "forceful" may be negatively related to work performance over and above the overall positive relation between masculinity and work performance; see Figure 6). For population Model 2A, there is no difference between isolated and distributed parceling because there is only one indicator that contains variance from another factor (i.e., the single offending item must be parceled with a set of correctly specified items). For Models 2B and 2C, I again compare isolated parceling (cross-loading items parceled together) with distributed parceling (cross-loading items parceled separately).

Results: Model 2A

Figure 7 (top left) shows the model-estimated *ab* for the item, parcel, and scale models as a function of the population cross-loading strength. When the cross-loading is 0, there is no misspecification in the item or parcel models. As the cross-loading becomes stronger, *ab* estimates in all models increase to account for the extra shared variance between *M* and an indicator of *X*. The resulting bias is more severe for the item model than the parcel model, especially as the cross-loading strength increases. When

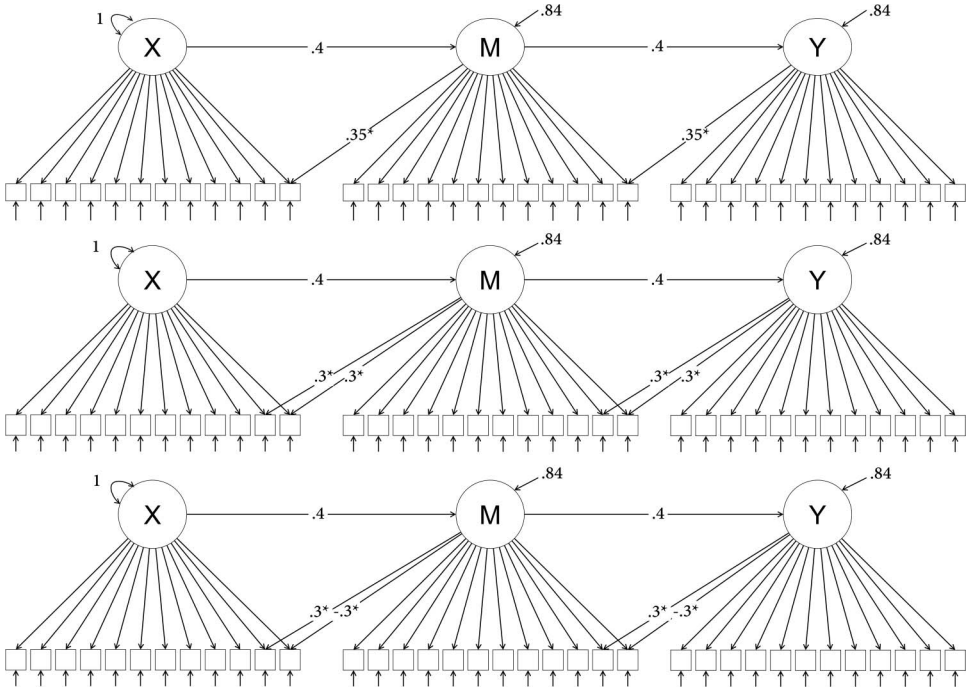

Figure 6. Population Models 2A (upper), 2B (middle), and 2C (lower). All primary factor loadings are .5. All manifest and latent variables have a total variance of 1. Asterisks denote population values that are varied continuously.

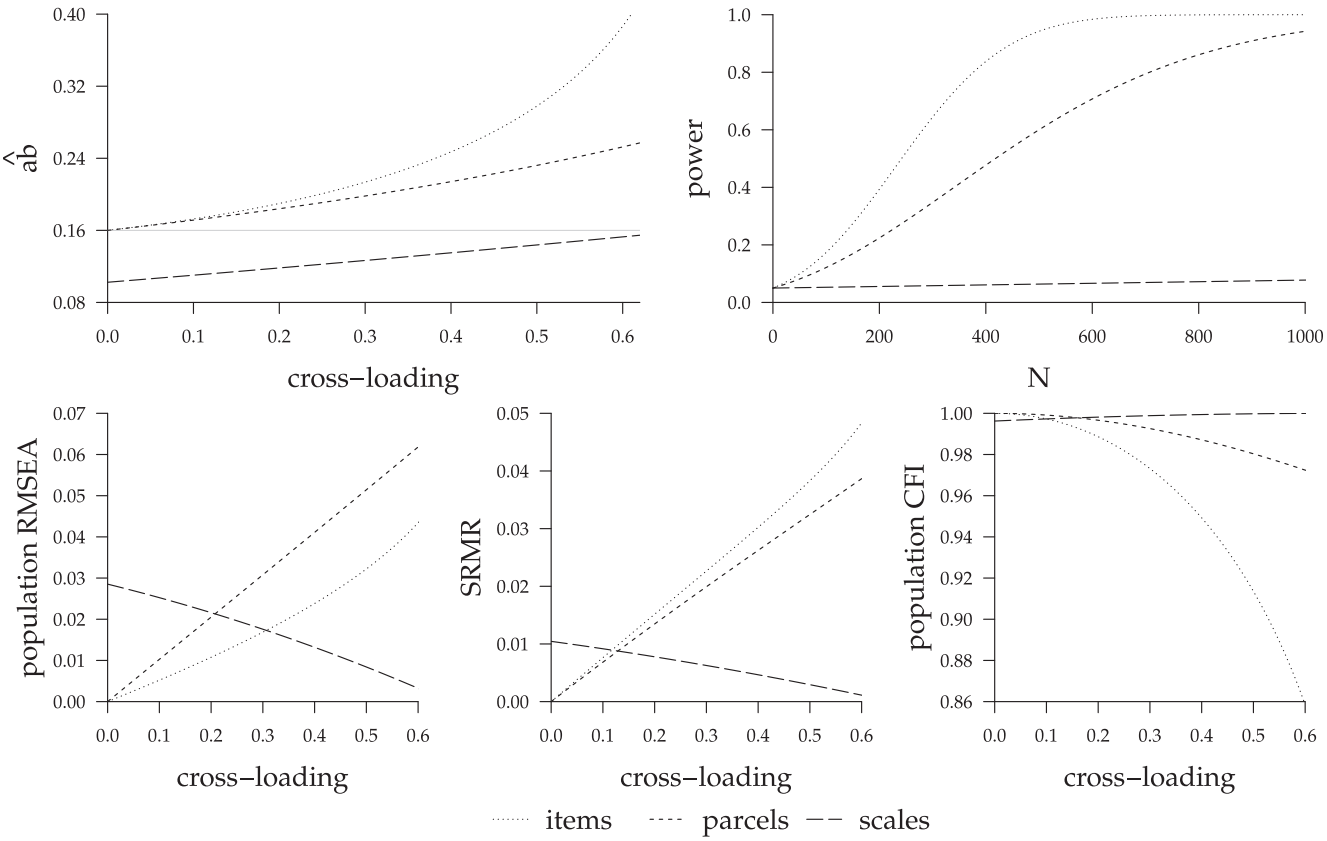

Figure 7. Model 2A results. Top left: Estimated values of *ab* as a function of the strength of the cross-loadings (*x*-axis) and parceling strategy (separate lines). The solid gray line at $ab = .16$ is the population value. Top right: Power to detect model misspecification by sample size (*x*-axis) and parceling strategy (separate lines) when the cross-loading value is .35. Bottom: Population fit indices as a function of the strength of the cross-loadings (*x*-axis) and parceling strategy (separate lines).

the cross-loading is .35, the item model *ab* is .229 (43% bias) and the parcel model *ab* is .206 (29% bias). Parameter bias in the scale model actually decreases as the misspecification becomes more severe, as two sources of bias compensate for each other: Measurement error creates negative bias, whereas the increasing cross-loading creates positive bias.

It is interesting that bias for all models is much more severe than in Study 1, despite approximately the same degree of misspecification (noncentrality) in the item models. For example, the item models in Study 1 showed 7% to 10% bias, compared with 43% bias here. Figure 7 (top right) shows the power to detect the misspecification when the cross-loading is .35. The item model has the highest power to detect the misspecification, reaching 80% power at $N = 376$, whereas the parcel model requires $N = 707$ to reach 80% power. As in Study 1, misfit in the scale model stems from attenuation due to measurement error rather than from the imposed measurement model misspecification, and power to detect this misspecification is very low.

The bottom row of Figure 7 displays fit index performance as a function of the cross-loading strength. Whereas SRMR and CFI are most sensitive to misspecification in the item model, RMSEA is most sensitive to misspecification in the parcel model. The scale model fits better by all three fit indices as the cross-loading becomes more severe; this result corresponds to the decrease in parameter bias for the scale model with increasing cross-loading strength.

Results: Model 2B

Model 2B includes two indicators with positive cross-loadings, making it possible to compare isolated parcels (parceling the two misbehaving items together) with distributed parcels (parceling them separately). Figure 8 (top) shows *ab* and power for the models based on items, isolated and distributed parcels, and scales as a function of the strength of cross-loadings (all four cross-loadings were varied simultaneously). These results follow the same trend as Model 2A, though the misspecification is greater, resulting in more bias and higher power overall. Distributed parcels minimize the bias in *ab*: When the cross-loadings are .3, *ab* for distributed parcels is .230 (44% bias), compared with .249 for isolated parcels (55% bias), .266 for items (66% bias), and .15 for scales (−5% bias; as in Model 2A, low bias is a result of the negative bias caused by measurement error being canceled out by upward bias due to cross-loadings).

Power curves (Figure 8, top right) reveal an interesting divergence: Although isolated parcels lead to less bias than items, they

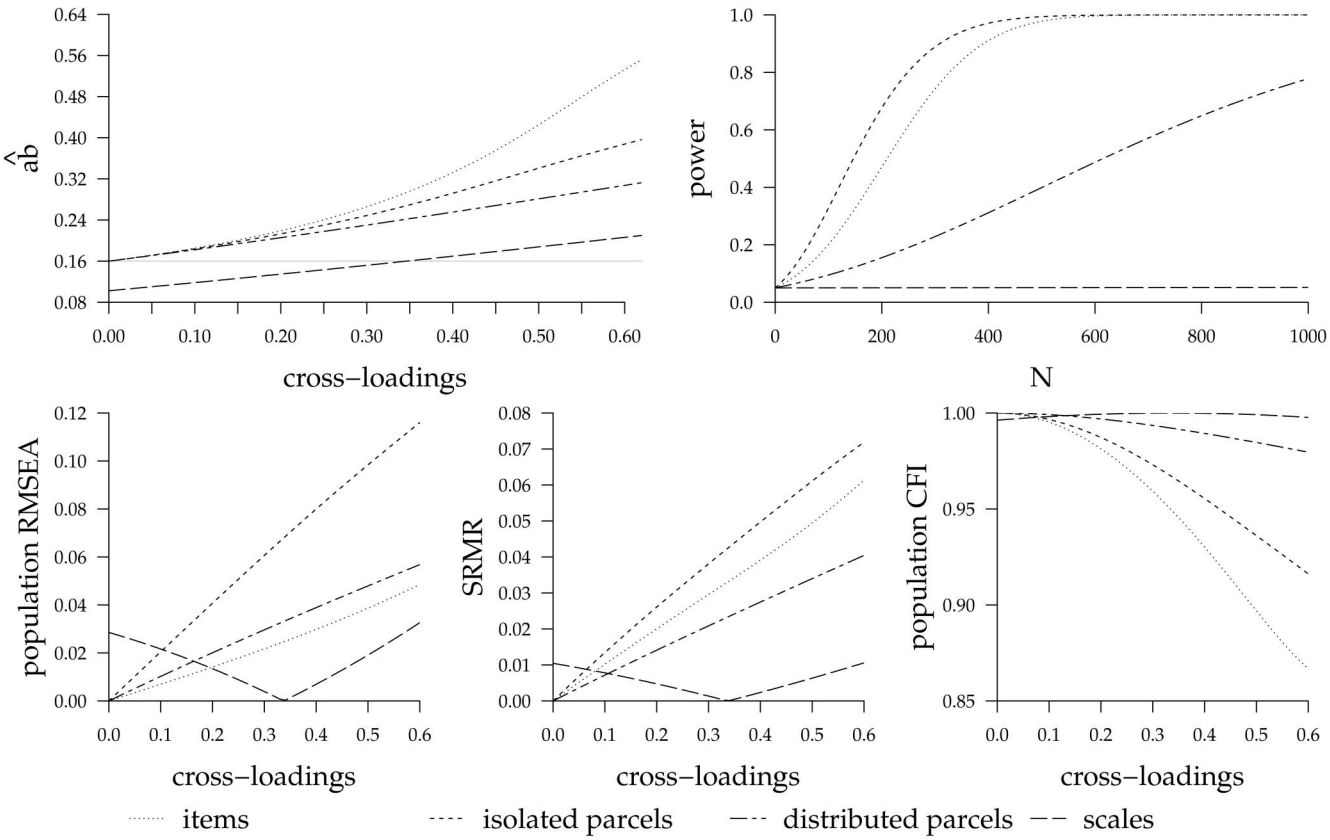

Figure 8. Model 2B results. Top left: Estimated values of *ab* as a function of the strength of the cross-loadings (*x*-axis) and parceling strategy (separate lines). The solid gray line at *ab* = .16 is the population value. Top right: Power to detect model misspecification by sample size (*x*-axis) and parceling strategy (separate lines) when the two cross-loading values are both .3. Bottom: Population fit indices as a function of the strength of the cross-loadings (*x*-axis) and parceling strategy (separate lines).

also lead to higher power to detect the misspecification. Distributed parcels, in contrast, have even less bias but substantially less power. When the cross-loadings are .3, the model based on items reaches 80% power at $N = 327$, isolated parcels require $N = 248$, and distributed parcels require $N = 1{,}041$. These results are consistent with Bandalos (2002, Study 2), in which an unmodeled method factor affected two constructs. As with the cross-loadings present here, that situation also results in indicators of two factors sharing extra variance.

The bottom row of Figure 8 displays fit indices as a function of cross-loading strength. CFI is, again, most sensitive to misfit in the item model. RMSEA and SRMR, in contrast, are both most sensitive to misfit in isolated parcels. RMSEA is also more sensitive to misfit in distributed parcels than in items. As in Model 2A, misfit in the scale model corresponds to its parameter bias: The two sources of bias perfectly cancel each other out when the cross-loadings are .33, and at this value, fit is perfect according to all three fit indices.

When four 3-item parcels are used, both isolated and distributed parcels result in slightly higher bias and higher power than their three 4-item parcel counterpart models (results not shown).

Results: Model 2C

In Model 2C, the valence of one cross-loading was flipped. The effect of opposite loadings is to reduce the bias in *ab*, for items as well as parcels, compared with two positive cross-loadings. Because one item has a positive cross-loading and another item has a negative cross-loading of the same strength, when the two items are combined into a parcel or scale, these variances fully cancel each other out. As a result, the isolated parcels strategy results in no bias, and scales result in only their baseline level of bias due to measurement error, regardless of the cross-loading strength (see Figure 9, top left). When the two items are entered into a measurement model as separate indicators (whether as items or distributed parcels), the bias does not completely cancel out. The bias pattern resulting from items and distributed parcels is the same as it was when both cross-loadings were positive: Items result in most bias (*ab* = .196; 23% bias) and distributed parcels result in about half as much bias (*ab* = .174; 9% bias).

Interestingly, power to detect the misspecification is almost equal for items (80% power at $N = 285$) and distributed parcels (80% power at $N = 324$; see Figure 9, top right), despite much greater bias in the item model. The bottom row of Figure 9 shows

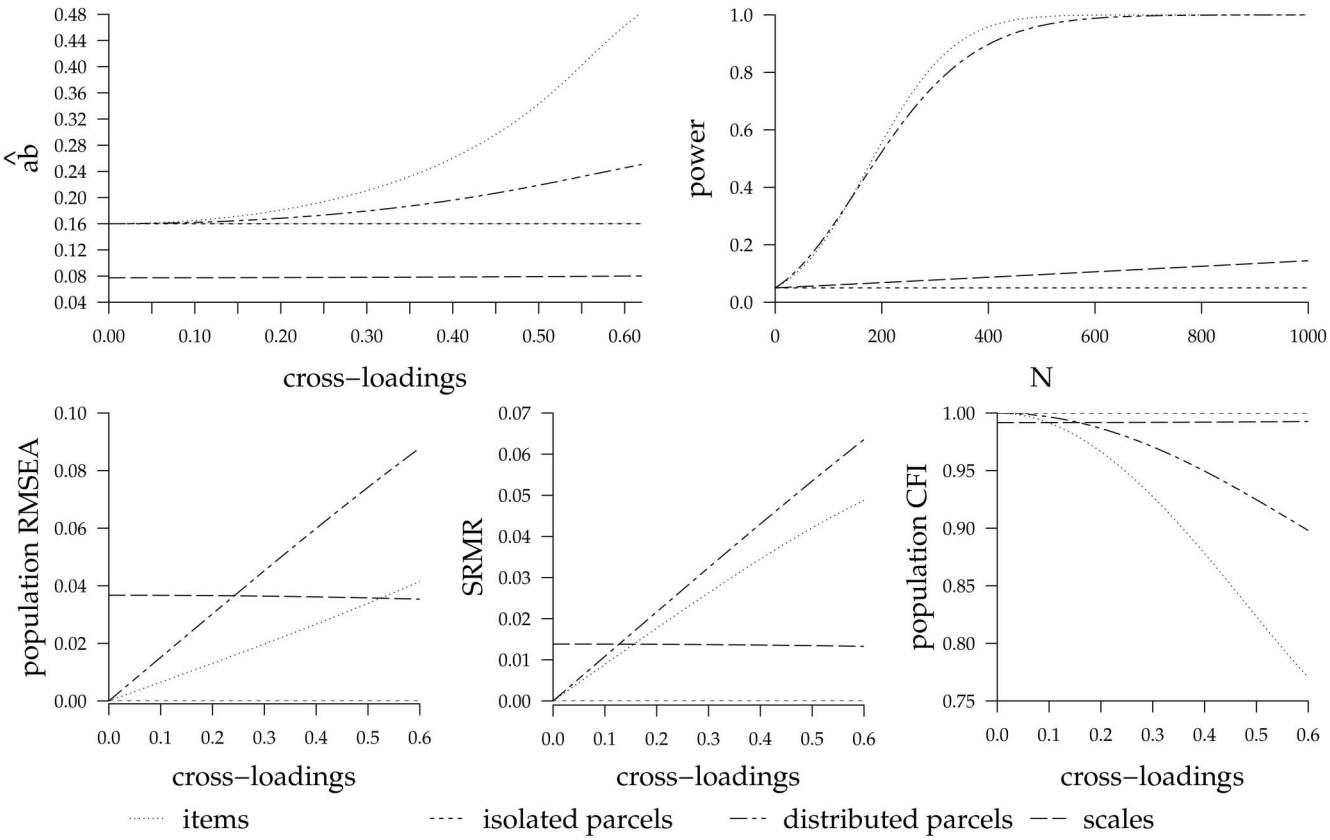

Figure 9. Model 2C results. Top left: Estimated values of *ab* as a function of the strength of the cross-loadings (*x*-axis) and parceling strategy (separate lines). The solid gray line at $ab = .16$ is the population value. Top right: Power to detect model misspecification by sample size (*x*-axis) and parceling strategy (separate lines) when the cross-loading values are .3 and −.3. Bottom: Population fit indices as a function of the strength of the cross-loadings (*x*-axis) and parceling strategy (separate lines).

population fit indices as a function of cross-loading strength: RMSEA and SRMR are most sensitive to the misspecification with distributed parcels, and CFI is again most sensitive to misfit in the item model.

Study 3: Structural Model Misspecification

Studies 1 and 2 showed that parceling can reduce parameter bias and misfit due to misspecifications within or across measurement models. Study 3 addresses the effect of parceling on structural model misspecification. Methodologists have expressed concern that structural model misspecification will be harder to detect with parcels than items, simply because parceled models have a smaller covariance matrix to fit. For example, Bandalos (2002) argued that

> because the use of item parcels has the effect of reducing the number of data points that must be fit, solutions based on parcels will not yield as stringent a test of SEM models as would analyses based on the individual items. (p. 80)

It is notable that structural model misspecification has never been investigated in the context of parceling. After all, parcels are most often employed when researchers have come to the point of testing a theory in a full structural equation model, in an effort to reduce the size of the model (Bandalos & Finney, 2001). At this point in model testing, researchers are presumably most interested in detecting problems with the theory, that is, the structural model.

I consider two population models (see Figure 10). In population Model 3A, the regression coefficient from $M \rightarrow Y$ is reversed, so *Y* is actually a predictor of *M*, and *X* and *Y* are unrelated. Fitting the mediation model to this matrix will result in erroneous model-implied covariances between the indicators of *X* and those of *Y* (via the indirect effect), where none exist. Whereas in previous models the population structural regression coefficients were .4, in Model 3A, they were set to .65 to keep the degree of misfit for the item model comparable across all models. Because the true $M \rightarrow Y$ path is 0, the true value of $ab = .65 \times 0$ is also 0. In Model 3B, a direct effect of *X* on *Y* has been added to the mediation model. Here, the misspecification is due to the omission of the direct effect: Indicators of *M* and *Y* covary more than the model can account for.

Results: Model 3A

When there is no measurement model misspecification, there is no isolated/distributed distinction. Instead, figures present results based on three 4-item parcels per factor (as in Studies 1 and 2) as well as those based on four 3-item parcels per factor.

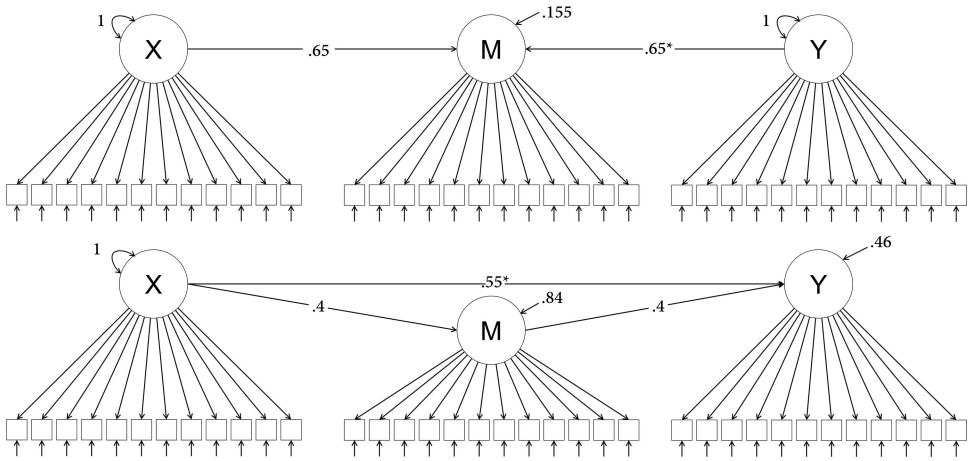

Figure 10. Population Models 3A (upper) and 3B (lower). All factor loadings are .5. All manifest and latent variables have a total variance of 1. Asterisks denote population values that are varied continuously.

Figure 11 (top left) displays estimates of *ab* as a function of the $Y \rightarrow M$ path (β_{MY}). When the value of the reversed path is 0, there is no misspecification for items or parcels. As the path becomes stronger, the bias increases. Parcels and items result in exactly the same bias. The indirect effect estimated from scale data is smaller than that of the latent variable models, because it is attenuated due to measurement error.

Figure 11 (top right) displays power to detect the structural misspecification for scales, three and four parcels, and items.

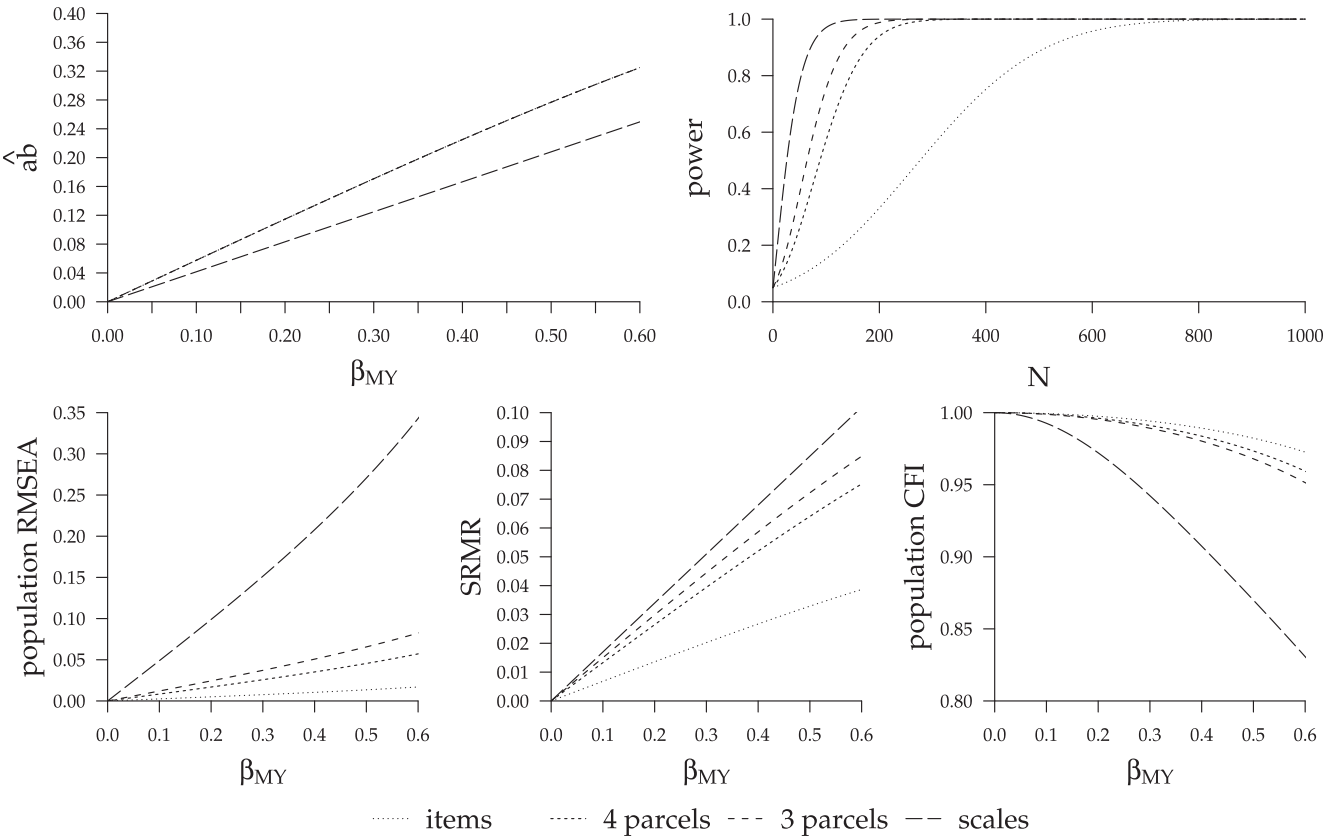

Figure 11. Model 3A results. Top left: Estimated values of *ab* as a function of the strength of the regression of *Y* on *M* (β_{MY}; *x*-axis) and parceling strategy (separate lines). The population value of *ab* is 0. Top right: Power to detect model misspecification by sample size (*x*-axis) and parceling strategy (separate lines) when $\beta_{MY} = .65$. Bottom: Population fit indices as a function of the strength of β_{MY} (*x*-axis) and parceling strategy (separate lines).

Scales and parcels result in considerably higher power to detect the misspecification than items. Although items require $N = 430$ to achieve 80% power, the same degree of power requires just $N = 108$ with parcels and $N = 53$ with scales. The reason for this difference has to do with degrees of freedom. The degree of misspecification is the same for items and three or four parcels—the lavaan estimates of the noncentrality parameter for each of these models differed by at most 0.6%—but the item model has many more degrees of freedom (592) than the four-parcel model (52 *df*), the three-parcel model (25 *df*), and the scale model (1 *df*). Because the test statistic is compared with a chi-square distribution that is centered on the model degrees of freedom, the item model requires a higher degree of misfit to reject the model.

Differences in model fit are even more noticeable in the RMSEA, which is adjusted for degrees of freedom (Figure 11, bottom). For example, when $\beta_{MY} = .65$, the population RMSEA of the item model is just .019, compared with .063 for four parcels and .092 for three parcels. Similarly, the SRMR is twice as high for three parcels as items (.080 for four parcels and .091 for three parcels compared with .041 for items), and the population CFI is substantially lower (.951 for four parcels and .941 for three parcels compared with .966 for items). All fit indices reveal even greater misspecification for scales, where the population SRMR = .110, RMSEA = .384, and CFI = .810, due to the low *df* of the scale model as well as the misspecification due to measurement error.

Results: Model 3B

Whereas the structural misspecification in Model 3A resulted in the model accounting for more covariance between *X* and *Y* than was actually present, the misspecification in Model 3B results in the model accounting for *less* covariance between *X* and *Y* than is present. The results, however, follow exactly the same pattern as Model 3A.

Figure 12 (top left) displays estimates of *ab* as a function of the missing direct $X \rightarrow Y$ path. When the missing path is 0, the latent variable models are not misspecified. All latent variable models overestimate *ab* to the same degree as the strength of the missing path increases. The scale model produces more accurate estimates of the indirect effect as the direct effect increases because the positive bias cancels out the negative bias due to measurement error.

As with Model 3A, scales result in the highest power to detect the structural misspecification, followed by three then four parcels, followed by items. The bottom row of Figure 12 displays popula-

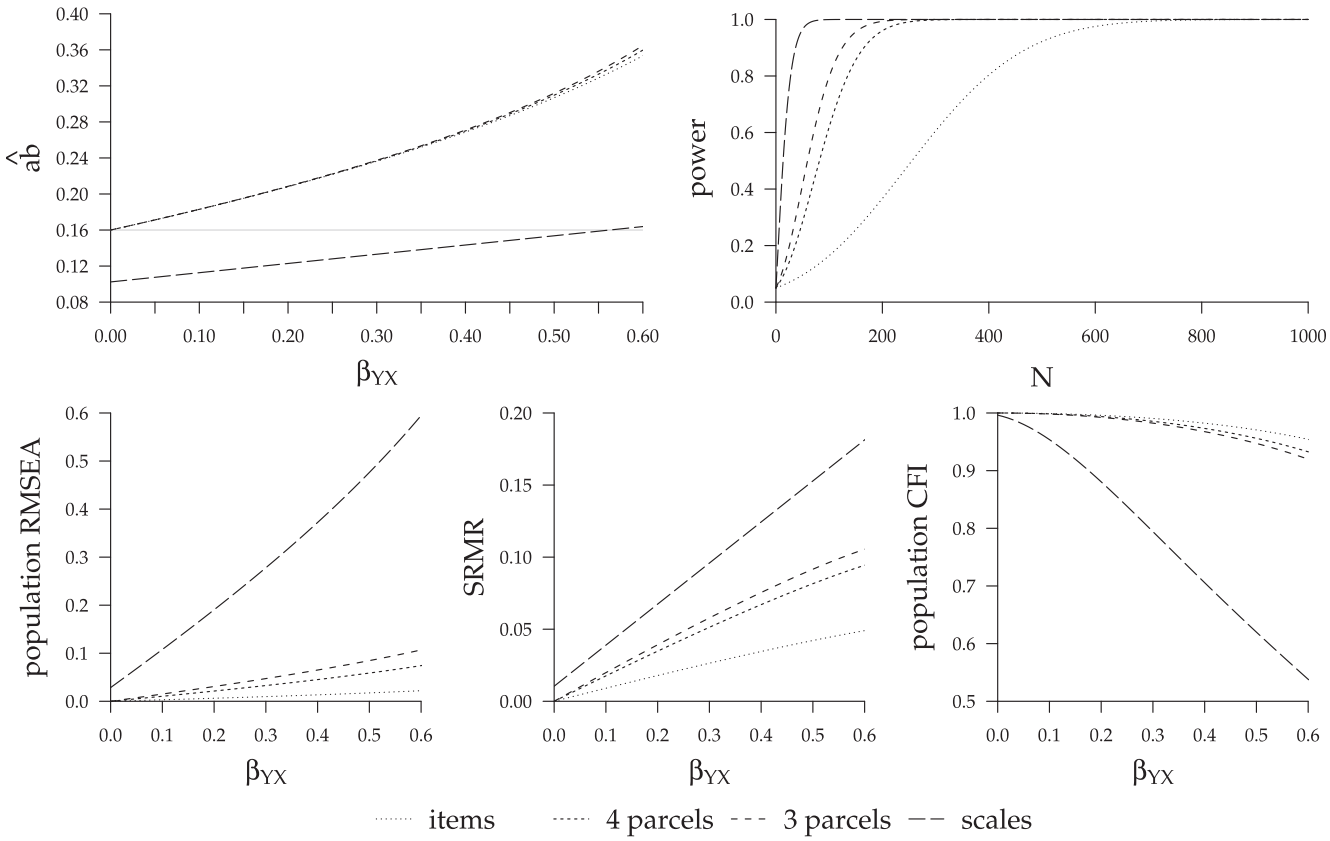

Figure 12. Model 3B results. Top left: Estimated values of *ab* as a function of the strength of the missing regression path (β_{YX}; *x*-axis) and parceling strategy (separate lines). The solid gray line at $ab = .16$ is the population value. Top right: Power to detect model misspecification by sample size (*x*-axis) and parceling strategy (separate lines) when $\beta_{YX} = .55$. Bottom: Population fit indices as a function of the strength of the missing regression path (*x*-axis) and parceling strategy (separate lines).

tion fit indices as a function of the missing direct path; these follow the same pattern as in Model 3A.

Discussion

Most psychological constructs are inherently complex, so to fully represent them requires heterogeneous items that do not behave according to the psychometric ideal of unidimensionality (Reise, Moore, & Haviland, 2010). If these items are modeled as though they are unidimensional indicators of a single latent construct, the measurement model will be misspecified and structural parameter estimates will be biased. One way to deal with measurement model misspecification is to model it explicitly via a bifactor or higher order factor model, a modeled method factor, or estimated cross-loadings or correlated residuals (Reise et al., 2010, 2013). These methods can clarify the multidimensional structure of a set of items and thereby illuminate the nature of the latent variable.

Parceling has been derided as a way to bury problems caused by poor items (e.g., Marsh, Lüdtke, Nagengast, Morin, & Von Davier, 2013) because a misspecified item-level measurement model can easily be transformed into a well-fitting parcel-level measurement model. Detractors argue that parceling produces latent variables that no longer represent what they were meant to represent, while still appearing to have good fit. The present results confirm that parceling can, in some situations, create good fit while preserving bias. However, they also show that, once the measurement properties of a set of items is well understood, strategic parceling can be used to create simple models with minimal bias in the structural model.

The present studies have also revealed a novel benefit of parceling that has nothing to do with measurement model misspecification: Parcels result in far greater power to detect structural model misspecifications compared with items. Thus, over and above any benefits that parcels bring to the measurement model, they may be worth using to achieve a stronger test of the structural model.

The Effect of Parcels on Bias

In line with previous research, when misspecified item-level models were parceled, structural parameter estimates often remained biased. Surprisingly, though, in almost no condition did a parceled model produce *more* bias in the structural coefficient than the item model did. With the exception of Model 1A, in which three 4-item isolated parcels resulted in slightly more bias than items for medium values of cor($X1$, $X2$), in every other model, both isolated and distributed parcels produced smaller bias in the mediated $X \rightarrow M \rightarrow Y$ path compared with items. This finding is consistent with Bandalos (2002), which is the only previous study that has directly compared parameter bias as a result of model misspecification in items versus parcels.

In some situations, one parceling strategy was able to completely or almost completely eliminate bias. Study 1 considered two scenarios in which items contain multiple sources of systematic variance. When it is desirable for both sources of variance to be contained in the latent factor (e.g., when these were two facets of the construct), then distributed parcels appropriately channel both sources of variance into the latent factor, producing a latent factor defined by both facets. When it is desirable for one source of variance to be relegated to residual variance (e.g., when a subset of items is affected by method variance), then isolated parcels achieve this goal by allocating all method variance to a single parcel, in which it ends up in that parcel's residual variance rather than in the factor. Thus, if the measurement properties of a set of items are known, a researcher can allocate items to parcels in such a way that factor-relevant variance is shared among parcels and factor-irrelevant variance is not.

For example, a researcher studying the effect of perfectionism on school achievement may not be interested in the underlying six-dimensional structure of perfectionism (Frost, Marten, Lahart, & Rosenblate, 1990). In creating parcels out of the Multidimensional Perfectionism Scale (Frost et al., 1990), one strategy would be to create isolated parcels that each reflect one facet of perfectionism. The resulting latent construct would behave like a higher order factor model, capturing only what is shared among the six facets, but none of the unique aspects of each facet (Coffman & MacCallum, 2005; Graham & Tatterson, 2000; Hagtvet & Nasser, 2004). In contrast, a distributed parcels strategy would allow the latent factor to reflect all the reliable variance in the construct, including unique and common aspects of each perfectionism dimension. Neither of these schemes is inherently better. Which one is chosen depends on whether the researcher intends to construe perfectionism as a higher order unidimensional construct (i.e., that part of perfectionism that is independent of its facets) or as a multidimensional construct.

Study 2 represents one of the first examinations of the effects of parceling on cross-factor measurement misspecification. When a cross-loading was present in the population, any parceling strategy reduced its effect and thereby reduced bias in the structural coefficient. Thus, one recommendation from this study is that any parceling strategy may be beneficial to reducing the effect of cross-factor measurement misspecification. When more is known about the structure of the misspecification, however, it may be possible to choose a parceling scheme that reduces the bias even further. In the case of two cross-loadings of the same valence, allocating these to separate parcels produces less bias. In the case of two cross-loadings of opposite valence, allocating these to the same parcel produces less bias. These results suggest that the accuracy of structural parameter estimates can be maximized by creating parcels based on a multifactor measurement model rather than a series of single-factor models. The multifactor measurement model may reveal substantial cross-loadings that can be taken into account to minimize their effect on the structural model. In contrast, parceling schemes based on single-factor CFA models cannot possibly account for cross-factor measurement model misspecification.

Finally, Study 3 confirmed that when misfit in the item model is entirely due to structural model misspecification, parceling does not introduce additional bias.

These studies confirmed the well-known finding that manifest variable models based on scales are susceptible to bias as a result of measurement error. In the absence of other misspecification, measurement error attenuated the mediated path in the scale model by 36%. When other measurement model misspecifications are present, their effects either compound or counteract this bias. These results affirm recommendations by Cole and Preacher (2014) and Coffman and MacCallum (2005) that latent variable

models be used instead of scales (but whereas Cole & Preacher recommended the full item-level model be used, Coffman & MacCallum recommended parcels).

The Effect of Parcels on Power to Detect Misspecification

Power to detect misfit by the chi-square test of exact fit depends on the misspecification. This article examined seven misspecified models, each with roughly equal power in the item-level model. Study 1 revealed that within-factor misspecifications, including a single factor fit to an underlying two-factor structure and a set of items with shared method variance, result in substantially lower power for parceled models than items. More optimistically for parcels, Study 2 revealed that power to detect cross-loadings ranges from somewhat lower for parcels, in the case of a single cross-loading, to slightly higher for parcels, in the case of multiple cross-loadings parceled together. Much more optimistically for parcels, Study 3 revealed that power to detect structural model misspecifications, including a misdirected path and a missing path, is considerably higher for parcels than items. The behavior of four parcels fell in between that of items and that of three parcels per factor, but was much closer to three parcels.

A manifest variable model fit to scale-level composites had virtually no power to detect a measurement model misspecification, but very high power to detect structural misspecifications. The finding that manifest variable models fit to scale data result in substantially higher power to detect structural misspecification is surprising based on current recommended practice. For example, Cole and Preacher (2014) found that as unreliability in a scale increases, power to detect model misspecification decreases substantially. They recommended that latent variable models be used instead of scales, and that these models be based on as many items as possible, on the basis of previous research suggesting that more indicators lead to more stable SEM solutions (Marsh et al., 1998; Mulaik, 2009). The present results suggest, however, that latent variable modeling does not improve power to detect structural model misspecification: In a given data set, using a latent variable model instead of scales leads to severely decreased power to detect the misspecification. Parcels provide a partial solution to this problem—by reducing the size of the model, parcels increase power to detect structural model misspecification compared with items, though power is still lower than for scales. Thus, a better recommendation, to maximize both accuracy and power, is to use latent variable models based on a small number of parcels.

The Effect of Parcels on Model Fit Indices

It is common practice for the fit of any structural equation model to be evaluated using not only the chi-square test of exact fit but also any of a number of popular fit indices. I examined RMSEA, SRMR, and Bentler's comparative fit coefficient, which is consistently estimated by CFI, NFI, and IFI. The results from these investigations reveal optimal strategies for detecting misfit in parceled models.

Population RMSEA is the square root of the estimated chi-square noncentrality parameter (which is used to obtain the chi-square test statistic) divided by the model degrees of freedom. Given the same degree of misfit, a model with more degrees of freedom will thus result in better fit. When the misspecification was in the measurement model (Studies 1 and 2), the noncentrality parameter was always substantially higher for the item model than the parceled models, which typically led to higher power for the item model. However, when these noncentrality parameters were divided by the models' respective degrees of freedom (592 *df* for the item model vs. 25 *df* for the parceled model), the result was frequently a higher RMSEA for one or both of the parceled models. When the misspecification was in the structural model (Study 3), the noncentrality parameters for item and parceled models were almost identical, resulting in substantially higher RMSEAs for the parceled models. These findings suggest that RMSEA may be a particularly sensitive index for discovering model misspecification in parceled models.

Similar to RMSEA, SRMR includes a type of parsimony correction because it indexes the average standardized residual over a smaller covariance matrix for parcels than for items. SRMR also showed promising performance for detecting misspecification in parceled models, producing the highest SRMR for parcels in almost all of the situations that RMSEA did. Of the two, RMSEA showed slightly higher sensitivity to misfit in parceled models.

The population version of the CFI, NFI, and IFI always produced substantially worse fit for items than parcels under any measurement model misspecification. Only under structural model misspecification, when power was already much higher for parceled models, did this index show (slightly) worse fit for parcels than items. These findings suggest that CFI, NFI, and IFI should not be used to judge the fit of a parceled model.

No fit index was able to detect measurement model misspecification in the scale model. The scale-based RMSEA, SRMR, and CFI were unaffected by the degree of measurement model misspecification—although in some situations these indices were highest for scales, they did not display worse fit as the misspecification became worse; in fact, they often suggested better fit with greater measurement model misspecification. In contrast, when the misspecification was at the structural level, all indices displayed substantially worse fit for scales than parcels or items.

Recommendations

Five recommendations can be distilled from the present set of findings:

1. The full item-level measurement model (i.e., a CFA model including all constructs) should be fit before deciding whether and how to allocate items to parcels. Not only is this is the most powerful way to detect measurement model misspecification—it may be used to identify sources of misfit (e.g., shared variance among items within and across latent variables) that can be minimized using parcels.

2. An ideal parceling scheme can be selected based on a known item-level population model. As the item-level population model is never known, a plausible basis on which to allocate items to parcels is a strong understanding of the measurement properties of scale items (e.g., based on item content, previous research, or empirical tools designed for this purpose; Reise, Bonifay, & Haviland, 2013). Given this type of informa-

tion, a parceling scheme can be chosen to achieve the desired content of each latent variable. When indicators are not unidimensional, isolated and distributed parcels both serve to redefine the common factor: Isolated parcels combine items that share unique variance, removing that variance from the latent factor and relegating it to residual variance. Distributed parcels allow all variance that is shared by any items to become part of the latent factor. The decision to use isolated or distributed parcels is likely to affect structural parameter estimates.

3. Any parceling scheme is likely (though not guaranteed) to reduce bias in structural parameters that is due to measurement model misspecification. Moreover, any parceling scheme will result in substantially higher power to test the fit of a structural model compared with items. As such, once the goal is to estimate structural parameters and/or to test a theory instantiated in a structural model, parceling is recommended.

4. When there is misfit across factors, such as a single cross-loading, any parceling scheme will substantially reduce its effects. If more than one item cross-loads on the same factor, allocating these to separate parcels will minimize their effects; if multiple items have cross-loadings on the same factor in opposite directions, allocating these to the same parcel can cancel out the effects. The present study has *not* examined the effects of parceling decisions based on observed sample misfit—for example, a cross-loading in a particular sample that may or may not exist at the population level.

5. To evaluate the fit of a parceled model, RMSEA and SRMR are a better bet than CFI or the chi-square test. Fit based on CFI and chi-square is virtually guaranteed to improve after parceling, *whether or not bias is reduced*. In contrast, RMSEA and SRMR may get worse after parceling, *even when bias is reduced.* An increase in RMSEA or SRMR after parceling does not indicate that the item model is better than the parceled model—it indicates that there is still misspecification in the model that may be causing bias.

Limitations and Future Directions

The present studies did not consider precision of parameter estimates under item, parcel, and scale models. Marsh et al. (1998) reported that the variability of structural parameter estimates across repeated samples is identical whether items or any number of parcels are modeled. Ledgerwood and Shrout (2011) reported that latent variable models produce higher variability in structural parameter estimates than do scale models: Their estimated indirect (mediation) path ranged from 70% more efficient (small effect size, high scale reliability) to 270% more efficient (large effect size, low scale reliability) than that estimated with a latent variable model. These findings are confirmed by examining the asymptotic standard errors of estimates in the present study (results not presented), which reveals that the variability of structural parameters was identical for items and parcels, but much lower for scales (e.g., with no model misspecification, the indirect path coefficient estimated with scales was 63% more efficient than the one estimated with items or parcels). The higher precision in scale models translates to higher power to detect nonzero structural coefficients (e.g., by a Wald test), though this benefit must be weighed against the loss of accuracy in scale models compared with latent variable models (Ledgerwood & Shrout, 2011).

Although the asymptotic results suggest that parcel-based parameter estimates are no less efficient than item-based estimates, this equality may not hold up with sample data. Sterba and MacCallum (2010) revealed that the use of parcels introduces *allocation variability*, that is, variability across different allocations of items to parcels within a single sample data set. Allocation variability affects parameter estimates as well as fit indices, even when items are interchangeable in the population (Sterba, 2011; Sterba & MacCallum, 2010). Further research is needed to clarify to what extent the advantages of parceling may be outweighed by lower precision due to the latent variable model (compared with scales) and allocation variability (compared with items).

The study of population performance leads to precise recommendations of how and when to parcel based on exact knowledge of population conditions. When population conditions are unknown, it would be desirable to have empirical parceling schemes that reliably produce isolated and distributed parcels. The most commonly applied empirical approaches are meant to produce distributed parcels. For example, the item-to-construct balance approach (Little et al., 2002; also known as the *factorial algorithm* [Rogers & Schmitt, 2004]), combines items with high factor loadings with those with low loadings to create a set of maximally similar parcels. As items with high factor loadings are most likely to be unidimensional indicators of the latent factor, and those with low loadings are more likely to be affected by a secondary source of variance, combining items with high and low loadings is likely to create distributed parcels.

Producing isolated parcels may prove more difficult in practice. Rogers and Schmitt (2004) compared four empirical parceling schemes, of which two were meant to create isolated parcels and two were meant to create distributed parcels. They found that all four strategies resulted in largely indistinguishable results that resembled distributed parcels. However, their population model was a bifactor model with four secondary factors, out of which three parcels were formed. Even if an empirical strategy could have identified four unique sets of variance, these would have been forced into three quasi-isolated parcels. More research is clearly needed. Still, distributed parcels may be much easier to achieve in practice, if only because there are many more item-to-parcel allocations that will produce distributed parcels than isolated parcels. Although isolated parcels are ideal in certain situations, the present results suggest that distributed parcels tend to lead to acceptable levels of bias and power.

The present results suggest that parcels can and should be selected based on item properties revealed via multifactor CFAs. It is not clear, however, to what extent this recommendation should be followed on the basis of small sample data. For example, if model modification indices based on a multifactor item-level CFA suggest adding two cross-loadings, allocating items to parcels

based on this sample information may be a valid strategy, or, if these cross-loadings are the spurious result of sample fluctuations, it may introduce new bias. Research based on simulated sample data will be required to evaluate the reliability of this approach.

Conclusion

Although there is a great deal of debate in the methodological literature about the dangers of parcels, the present results suggest that researchers may benefit from using them. In most cases, bias in structural coefficients is reduced compared with item models. Though power to detect measurement model misspecification is frequently diminished, RMSEA and SRMR are more sensitive to this misspecification in parceled models. Most crucially, parceling vastly improves power to detect misspecification in the structural model. The goal of most research using SEM is to test the structural relations among constructs. Parcels help make that goal more attainable.

References

Alhija, F. N.-A., & 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

Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika, 49,* 155–173. http://dx.doi.org/10.1007/BF02294170

Babakus, E., Ferguson, C. E., Jr., & Jöreskog, K. G. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. *JMR, Journal of Marketing Research, 24,* 222–228. http://dx.doi.org/10.2307/3151512

Bagozzi, R. P., & Heatherton, T. F. (1994). A general approach to representing multifaceted personality constructs: Application to state self-esteem. *Structural Equation Modeling, 1,* 35–67. http://dx.doi.org/10.1080/10705519409539961

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

Bandalos, D. L. (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. L., & Finney, S. J. (2001). Item parceling issues in structural equation modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 269–296). Hillsdale, NJ: Erlbaum.

Bem, S. L. (1974). The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology, 42,* 155–162. http://dx.doi.org/10.1037/h0036215

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238–246. http://dx.doi.org/10.1037/0033-2909.107.2.238

Bentler, P. M. (2006). *EQS 6 Structural equations program manual*. Encino, CA: Multivariate Software.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588–606. http://dx.doi.org/10.1037/0033-2909.88.3.588

Bollen, K. A. (1989). *Structural equation models*. New York, NY: Wiley.

Boomsma, A. (1982). The robustness of LISREL against small sample sizes in factor analysis models. In K. G. Jöreskog & H. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction* (Part I, pp. 149–173). Amsterdam, the Netherlands: Elsevier.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. *Sage Focus Editions, 154,* 136.

Cattell, R. B. (1956). Validation and intensification of the sixteen personality factor questionnaire. *Journal of Clinical Psychology, 12,* 205–214. http://dx.doi.org/10.1002/1097-4679(195607)12:3<205::AID-JCLP2270120302>3.0.CO;2-0

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

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

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods, 19,* 300–315. http://dx.doi.org/10.1037/a0033805

Frost, R. O., Marten, P., Lahart, C., & Rosenblate, R. (1990). The dimensions of perfectionism. *Cognitive Therapy and Research, 14,* 449–468. http://dx.doi.org/10.1007/BF01172967

Graham, J. W., & Tatterson, J. W. (2000). *Creating parcels for multidimensional constructs in structural equation modeling (No. 00–41).* University Park, PA: The Methodology Center, The Pennsylvania State University.

Green, S. B., Akey, T. M., Fleming, K. K., Hershberger, S. L., & Marquis, J. G. (1997). Effect of the number of scale points on chi-square fit indices in confirmatory factor analysis. *Structural Equation Modeling, 4,* 108–120. http://dx.doi.org/10.1080/10705519709540064

Gribbons, B. C., & Hocevar, D. (1998). Levels of aggregation in higher level confirmatory factor analysis: Application for academic self-concept. *Structural Equation Modeling, 5,* 377–390. http://dx.doi.org/10.1080/10705519809540113

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

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

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Uppsala, Sweden: Scientific Software International.

Kim, S., & Hagtvet, K. A. (2003). The impact of misspecified item parceling on representing latent variables in covariance structure modeling: A simulation study. *Structural Equation Modeling, 10,* 101–127. http://dx.doi.org/10.1207/S15328007SEM1001_5

Kishton, J. M., & Widaman, K. F. (1994). Unidimensional versus domain representative parceling of questionnaire items: An empirical example. *Educational and Psychological Measurement, 54,* 757–765. http://dx.doi.org/10.1177/0013164494054003022

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

Ledgerwood, A., & Shrout, P. E. (2011). The trade-off between accuracy and precision in latent variable models of mediation processes. *Journal of Personality and Social Psychology, 101,* 1174–1188. http://dx.doi.org/10.1037/a0024776

Little, T. D. (2013). *Longitudinal structural equation modeling*. New York, NY: Guilford Press.

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., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1,* 130–149. http://dx.doi.org/10.1037/1082-989X.1.2.130

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4,* 84–99.

Marsh, H. W., Hau, K. T., Balla, J. R., & Grayson, D. (1998). Is more ever too much: The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33,* 181–220. http://dx.doi.org/10.1207/s15327906mbr3302_1

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

Mathieu, J. E., & Farr, J. L. (1991). Further evidence for the discriminant validity of measures of organizational commitment, job involvement, and job satisfaction. *Journal of Applied Psychology, 76,* 127–133. http://dx.doi.org/10.1037/0021-9010.76.1.127

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

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification, 6,* 97–103.

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

Moshagen, M. (2012). The model size effect in SEM: Inflated goodness-of-fit statistics are due to the size of the covariance matrix. *Structural Equation Modeling, 19,* 86–98. http://dx.doi.org/10.1080/10705511.2012.634724

Mulaik, S. A. (2009). Linear causal modeling with structural equations. Boca Raton, FL: CRC Press.

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38,* 171–189. http://dx.doi.org/10.1111/j.2044-8317.1985.tb00832.x

Muthén, B. O. (1993). Goodness of fit with categorical and other nonnormal variables. *SAGE Focus Editions, 154,* 205.

Muthén, B. O., du Toit, S. H., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. *Psychometrika, 75,* 1–45.

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 reactions to tests. *Applied Measurement in Education, 16,* 75–97. http://dx.doi.org/10.1207/S15324818AME1601_4

Poulin, F., & Boivin, M. (2000). Reactive and proactive aggression: Evidence of a two-factor model. *Psychological Assessment, 12,* 115–122. http://dx.doi.org/10.1037/1040-3590.12.2.115

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment, 95,* 129–140.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment, 92,* 544–559. http://dx.doi.org/10.1080/00223891.2010.496477

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement, 73,* 5–26. http://dx.doi.org/10.1177/0013164412449831

Rhemtulla, M., Brosseau-Liard, P. E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17,* 354–373. http://dx.doi.org/10.1037/a0029315

Rhemtulla, M., Savalei, V., & Little, T. D. (2014). On the asymptotic relative efficiency of planned missingness designs. *Psychometrika*. Advance online publication. http://dx.doi.org/10.1007/s11336-014-9422-0

Richardson, H. A., Simmering, M. J., & Sturman, M. C. (2009). A tale of three perspectives: Examining post hoc statistical techniques for detection and correction of common method variance. *Organizational Research Methods, 12,* 762–800. http://dx.doi.org/10.1177/1094428109332834

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

Rucker, D. D., Preacher, K. J., Tormala, Z. L., & Petty, R. E. (2011). Mediation analysis in social psychology: Current practices and new recommendations. *Social and Personality Psychology Compass, 5,* 359–371. http://dx.doi.org/10.1111/j.1751-9004.2011.00355.x

Saris, W. E., & Satorra, A. (1993). Power evaluations in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 181–204). London, UK: Sage.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.

Satorra, A., & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika, 50,* 83–90. http://dx.doi.org/10.1007/BF02294150

Savalei, V. (2012). The relationship between root mean square error of approximation and model misspecification in confirmatory factor analysis models. *Educational and Psychological Measurement, 72,* 910–932. http://dx.doi.org/10.1177/0013164412452564

Schaller, T. K., Patil, A., & Malhotra, N. K. (2015). Alternative techniques for assessing common method variance: An analysis of the theory of planned behavior research. *Organizational Research Methods, 18,* 177–206. http://dx.doi.org/10.1177/1094428114554398

Simmering, M. J., Fuller, C. M., Richardson, H. A., Ocal, Y., & Atinc, G. M. (2015). Marker variable choice, reporting, and interpretation in the detection of common method variance a review and demonstration. *Organizational Research Methods, 18,* 473–511. http://dx.doi.org/10.1177/1094428114560023

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25,* 173–180. http://dx.doi.org/10.1207/s15327906mbr2502_4

Steiger, J. H., & Lind, J. C. (1980, May). Statistically-based tests for the number of common factors. Paper presented at the Annual Spring Meeting of the Psychometric Society, Iowa City, IA.

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

Takahashi, T., & Nasser, F. (1996). The impact of using item parcels on ad hoc goodness of fit indices in confirmatory factor analysis: An empirical example. Paper presented at the annual meeting of the American Educational Research Association, New York. http://files.eric.ed.gov/fulltext/ED398279.pdf

Thompson, B., & Melancon, J. G. (1996). Using item "testlets"/"parcels" in confirmatory factor analysis: An example using the PPSDQ-78. http://files.eric.ed.gov/fulltext/ED404349.pdf

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

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

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. M., & 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

Received September 11, 2014
Revision received October 20, 2015
Accepted October 30, 2015 ■