2025 ■ Vol. 21 ■ no. 2

Exploring the Effect of Item Parceling Strategies and Number of Items per Parcel on Measurement Invariance Testing in Confirmatory Factor Analysis

Chunhua Cao [a] and Xinya Liang [b]

[a]University of Alabama, 520 Colonial Dr, Tuscaloosa, AL 35401

[b]University of Arkansas, 751 W Maple St., Fayetteville, AR, 72701

Abstract ■ Item parceling is common in structural equation modeling used in psychology, especially when the research interest is in the structural relationship between latent factors, but research on its performance in measurement invariance testing is scarce. This study examined the sensitivity of fit measures (LRT, RMSEA, CFI, AIC, BIC, and SaBIC) to measurement noninvariance with different numbers of items per parcel and parceling strategies using a simulation study. The design factors included the location of measurement noninvariance, the magnitude of measurement noninvariance, the proportion of items with measurement noninvariance, sample size, number of items per parcel, and parceling strategies. Results suggested that compared to no parceling, isolated parceling tended to increase the sensitivity of fit measures, especially when fewer parcels with more items per parcel were used; distributed parceling showed lower sensitivity, particularly when noninvariant items were evenly distributed and fewer parcels were used. The findings provide insights into the use parceling in examining the relationship between constructs. The paper concludes with practical implications and future research directions.

Keywords ■ measurement invariance, item parceling, distributed parceling, isolated parceling, fit measures.

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Reviewers ■ One anonymous reviewer

ccao6@ua.edu

10.20982/tqmp.21.2.p095

Introduction

When fitting a confirmatory factor analysis (CFA) with a large number of variables, it can be useful to create some parcels to be used as indicators for the latent factor to reduce model complexity. The use of item parcels in CFA and structural equation modeling (SEM) has become quite common in the past few decades (Bandalos, 2002). Parcels are created by aggregating the scores of two or more individual items (e.g., averaging, summing), and used as indicators of latent factors in place of individual items. Previous research showed that model fit deteriorated with a large number of observed variables even when the model was correctly specified (Kenny & McCoach, 2003; Marsh et al., 1998). Parceling into fewer and more reliable indicators can help reduce measurement errors and improve model fit (Bandalos, 2002, 2008; Hall et al., 1999; Landis et al., 2000; Nasser & Wisenbaker, 2003). In addition, parceling simplifies complex models with many items, helping to meet the statistical assumptions necessary for SEM, such as multivariate normality. It also provides a more manageable number of parameters to be estimated, which is particularly beneficial in situations with smaller sample sizes. By creating parcels that represent the broad construct, researchers can ensure that each composite score reflects the construct adequately, although it is essential to apply this method judiciously to avoid masking the true nature of the underlying structure of the data.

Despite the common use of items parcels in SEM, research on item parceling in the multi-group SEM context is limited, especially in measurement invariance testing (Lee & Whittaker, 2021). Measurement invariance holds when the measurement properties are identical regardless of group membership or time points. While multigroup CFA

(MGCFA) is a standard approach in testing measurement invariance, there has been scarcer research on measurement invariance testing in MGCFA with item parcels than with individual items. Meade and Kroustalis (2006) conducted a simulation study to examine the impact of items parcels on measurement invariance testing compared to individual items, and their results suggested that item parceling masked the detection of measurement invariance. However, the authors did not vary the number of items per parcel in this paper, and the number of items per parcel was fixed at four. When different numbers of items per parcel are used, the total number of parcels used as indicators of the latent factor subsequently differs. Nasser and Wisenbaker (2003) found better goodness-of-fit of CFA model using fewer parcels. Cao and Liang (2022) showed that the number of indicators per factor had an impact on the detection of measurement noninvariance in multi-group CFA. Thus, it is likely that different number of items per parcel has an impact on the detection of measurement noninvariance using item parceling. However, the impact of number of items per parcel in measurement invariance testing has not been examined in the literature. Another issue is the lack of systematic investigation into the effect of different parceling strategies such as the combination of items with different degrees of noninvariance. Two primary parceling strategies include the isolated strategy (putting the noninvariant items in one parcel) and distributed strategy (distribute the noninvariant items to different parcels). The choice of parceling strategy and the number of items per parcel can have profound implications for detecting measurement noninvariance.

There has been a long-standing debate on the use of item parceling in SEM despite its widespread application in psychology, education, business, sociology, and other social behavioral sciences. The proponents advocate for its use because of its statistical advantages. Item parcels tend to have better psychometric properties. For example, the aggregated parcels are more normally distributed than the individual items, especially when the items are categorical and nonnormal variables (Hau & Marsh, 2004; Little et al., 2002). Also, item parcels usually exhibit higher reliability and communality than the individual items (Little et al., 2013). The most notable benefit is that item parcels yield better model fit indices with a parsimonious measurement model (Bandalos, 2002; Rogers & Schmitt, 2004). On the other hand, some researchers oppose using item parcels because of potential problems, such as obfuscating the model misspecification (e.g., Marsh et al., 2013) and masking measurement noninvariance (Meade & Kroustalis, 2006). Other researchers suggest that item parceling can be useful in some circumstances but detrimental in other circumstances, and thus, should be used with caution (e.g., Little et al., 2013).

Measurement invariance testing using parcels is critical in multi-group SEM as measurement invariance is the prerequisite of comparing factor means and structural relationships between groups (Lee & Whittaker, 2021). The simulation study by Lee and Whittaker (2021) showed that the incorrect specification of measurement invariance using parcels resulted in biased estimates of factor means and structural relationship, as well as lower power for detecting the structural parameters. Similar to the testing measurement invariance using individual items, testing measurement invariance using item parcels involves a series of sequential models with different degrees of restriction of the measurement parameters. In the baseline model that only specifies identical pattern of items loading on the latent factors without any constraints of measurement parameters, the model fit statistics might not be good because of the large number of items (Kenny & McCoach, 2003). Applied researchers have chosen to use parcels instead of original items to achieve better model data fit in baseline model.

With regard to model fit evaluation, the likelihood ratio test (LRT) was the only fit measure used to evaluate the performance of parceling in Meade and Kroustalis (2006). Changes in goodness-of-fit indices, such as the comparative fit index (CFI) and root mean square error of approximation (RMSEA) have also been used frequently for measurement invariance testing in SEM, but not been extensively investigated in measurement invariance testing with item parceling. The use of information criteria (ICs), such as Akaike information criterion (AIC; Akaike, 1974), Bayesian information criterion (BIC; Schwarz, 1978), and sample size adjusted BIC (SaBIC; Sclove, 1987), has been relatively less common, but ICs have been used in detecting differential item functioning (DIF S.-H. Kim et al., 2019) or measurement noninvariance (e.g., Cao & Liang, 2022; Liang & Luo, 2020). Despite this, the ICs have not been investigated in measurement invariance testing using parcels.

In this study, we explored the impact of parceling strategies and number of items per parcel on the performance of fit measures (LRT, CFI, RMSEA, AIC, BIC, and SaBIC) in detecting measurement noninvariance. To achieve these goals, we conducted a Monte Carlo simulation study to examine the effects of parceling on measurement invariance testing using MGCFA under a variety of research scenarios.

Literature Review

Item Parceling

Item parceling is based on heuristic consideration rather than theoretical consideration (Bandalos, 2008). It has been attracting researchers' attention in the past few decades because of the benefits it demonstrated. In a complex SEM model with a large number of variables, using item parcels yields a more parsimonious model with an optimal indicator- to-sample size ratio, especially when the sample size is not big enough. In this way, the SEM model with parcels is more likely to achieve convergence and stable estimates of the parameters. The often-cited advantage of item parceling is its improved model fit compared to the SEM model using the unparcelled individual items, which has been demonstrated in several previous simulation studies (Bandalos, 2002, 2008; Hall et al., 1999; Landis et al., 2000). However, even in misspecified SEM models, item parceling resulted in excellent fit, masking model misspecification. This calls the use of item parceling into question in SEM (Bandalos, 2002, 2008; Marsh et al., 2013), and its efficacy and applications have been the topic of debate. Moreover, items parceled together should be unidimensional, otherwise severe bias in structural parameters estimates was observed (Bandalos, 2008).

Another psychometric advantage of using item parceling in SEM is that the aggregated parcels usually result in more normally distributed indicators than the original individual items, especially when the items are nonnormal or categorical. Parcels that are more continuous and conform to normality better meet the assumption of normality of the data, producing more stable model fit indices and parameter estimates using maximum likelihood estimator (ML) (Hall et al., 1999; Little et al., 2013). However, if nonnormality is the only concern, there is no need to resort to item parceling since the estimator of weighted least squares with mean and variance adjusted (WLSMV) has good performance in detecting model misspecifications (Bandalos, 2008). In addition, properly aggregated item parcels often exhibit higher reliability and communality than the original items, underscoring the critical role of strategic parceling.

There are two major strategies for parceling based on how the items are combined. The first strategy is isolated parceling, in which the items with common source of variance are combined into parcels. The other strategy is distributed parceling, in which, items sharing the source of variance are put into different parcels (Bandalos, 2002). When the SEM model was misspecified, distributed parceling yielded lower rejection rates than isolated parceling and unparcelled conditions (Bandalos, 2002). In other words, distributed parceling could mask model misspecification to a larger degree compared to isolated parceling. The global fit index of RMSEA was substantially lower and CFI was higher (both indicating better model fit) in the distributed than the isolated strategy. Moreover, the results in Bandalos' study showed that more items per parcel resulted in lower rejection rates of the misspecified model. That is to say, the number of items per parcel played a role in detecting model misspecification in SEM. In the context of measurement invariance testing, distributed parceling can be formed by distributing noninvariant items to different parcels, whereas isolated parceling is created by combining the noninvariant items in the same parcel(s). Previous research showed that within the same sample, different random item-to-parcel allocations resulted in variability in structural parameter estimates and model fit indices (Sterba & MacCallum, 2010; Sterba, 2011). However, isolated and distributed approaches represent the two extremes of possible item parceling strategies, and any other methods to create parcels (e.g., not complete isolated or distributed parceling strategies) are expected to be in-between the two parcel-allocation variability (Sterba, 2011).

In summary, parceling is commonly used in SEM for various reasons, but it should not be used as a magical alternative to the original items (Lee & Whittaker, 2021). Instead, caution should be exercised in applying parceling techniques to make sure that the psychometric characteristics of individual items are well understood. Little et al. (2013) pointed out that “parcels can have terrible effects under certain circumstances and brilliant effects under other circumstances” (p. 3). Previous research showed the problems with item parceling for measurement invariance testing with four items per parcel (Meade & Kroustalis, 2006), but it might not have provided a complete picture of the performance of parceling in the context of measurement invariance testing. This study aims to explore the efficacy of item parcels using different parceling strategies in measurement invariance testing evaluated by various fit measures.

Measurement Invariance in Multi-group CFA

In MGCFA, there could be multiple observed groups in the population (e.g., gender groups, control and treatment groups). Measurement invariance implies that the relationships between the observed indicators and the latent factors are invariant for members from different observed groups. With continuous indicators or item parcels, the linear relationship between the observed scores and the underlying latent factors can be expressed as:

$$\mathbf{y}_{jg} = \nu_g + \mathbf{\Lambda}_g \eta_{jg} + \delta_{jg}. \tag{1}$$

For person j in group g, $\mathbf{y}_{jg}$ represents a $p \times 1$ vector of observed scores, ν_g is a $p \times 1$ vector of item intercepts for group g, $\mathbf{\Lambda}_g$ is a $p \times m$ matrix of item loadings, η_{jg} is a $m \times 1$ vector of latent factor scores, δ_{jg} is a $p \times 1$ vector of measurement errors, and its variance and covariance matrix is denoted as $\mathbf{\Theta}_g$ assumed to be diagonal. Testing measurement invariance in MGCFA is typically conducted by running a set of more restrictive models sequentially (Mered-

Table 1 ■ Summary of the Features of Configural, Metric, and Scalar Invariance Testing

Factor Loadings	Intercepts	Residual Variances	Purpose / Interpretation
Configural Invariance			
Free	Free	Free	Tests whether the same factor structure holds across groups (i.e., same pattern of loadings).
Metric Invariance			
Equal	Free	Free	Tests whether the factor loadings are equal across groups, indicating comparable construct meanings.
Scalar Invariance			
Equal	Equal	Free	Tests whether both loadings and item intercepts are equal, allowing for meaningful comparisons of latent means.

Note. *Free* means not constrained whereas *Equal* means constrained.

ith, 1993; Millsap, 2011). The first model tested is the least restricted configural invariance model, testing equal factor structures of which items loading on which factors across the groups, and no constraints are imposed on the factor loadings, item intercepts, and residual variances. If configural invariance holds, metric invariance testing is conducted, in which the factor loadings are constrained equal across groups ($\mathbf{\Lambda}_{\mathrm{g}} = \mathbf{\Lambda}$). When factor loadings are constrained equal across groups, a common practice is to freely estimate the factor variances and covariances in all groups except for the reference group. When metric invariance is tenable, scalar invariance can be tested, in which item intercepts are constrained to be equal across groups in addition to factor loadings ($\nu_{\mathbf{g}} = \nu$ and $\mathbf{\Lambda}_{\mathrm{g}} = \mathbf{\Lambda}$). Scalar invariance testing is important since it is the prerequisite for comparing latent factor means (Meredith, 1993). When intercepts are constrained to be equal across groups, latent factor means are freely estimated except for those in the reference group, which are typically constrained to be zero. If scalar invariance holds, strict invariance is conducted, in which the residual variances across groups are constrained to be equal in addition to equal factor loadings and items intercepts ($\Theta_g = \Theta$, $\nu_{\mathbf{g}} = \nu$, and $\mathbf{\Lambda}_{\mathbf{g}} = \mathbf{\Lambda}$). In this study, we limited measurement invariance testing to metric and scalar invariance testing, as scalar invariance is often considered to be sufficient for cross-group comparisons (Little, 1997; Widaman & Reise, 1997). A summary of the features of three levels of measurement invariance testing is provided in Table 1.

Several fit measures have been used in evaluating measurement invariance testing. First, LRT, or χ^2 difference testing, can be used to compare the fit of nested models in measurement invariance testing in SEM. The χ^2 of the more restrictive model is compared to that of the less restrictive model, and the χ^2 difference between the two competing models follows a χ^2 distribution with degrees of freedom equal to the difference in model parameters. The χ^2 difference can be tested using a χ^2 significance testing with its associated degrees of freedom. The null hypothesis of the χ^2 difference testing is that the more restrictive model fits the data as well as the less constrained model. If the χ^2 difference testing is not significant, the more constrained model is selected for being parsimonious because the two models fit the data equally well. On the other hand, a significant χ^2 difference testing indicates that the less constrained model fits statistically better than the more constrained model. In the context of measurement invariance testing, a nonsignificant χ^2 difference testing indicates that measurement invariance holds. Second, the goodness-of-fit indexes of CFI and RMSEA have been commonly used by both methodologists and applied researchers in evaluating measurement invariance testing because of their performance in evaluating both overall model fit and comparative model fit (e.g., Chen, 2007; Cheung & Rensvold, 2001). Cheung and Rensvold (2001) recommended a less than .01 decrease in comparative fit index (CFI) of the more restrictive model to support for measurement invariance. Chen (2007) suggested a less than .01 decrease in CFI in the more constrained model, supplemented by an increase of less than .015 in root mean square error of approximation (RMSEA) in the more constrained model to support for measurement invariance. The ICs, such as AIC, BIC, and SaBIC have also been used for comparing competing models. They are computed using the log likelihood of a fitted model, and each of them includes a different penalty term as a function of the number of parameters, sample size, or both. ICs can be used to compare the less and more constrained models in measurement invariance testing, with a lower value associated with a better model. Compared to AIC, BIC imposes a harsher penalty for the number of parameters estimated in the models, favoring the simpler model (e.g., Vrieze, 2012). SaBIC imposes a less harsh penalty for model complexity compared to BIC by including an adjusted term of sample size. The performance of the ICs in measurement invariance testing using item parcels remains unknown.

Purpose of This Study

As shown in the literature review above, item parceling has been studied and applied extensively in SEM. However, its efficacy in measurement invariance testing has not been examined comprehensively. To examine the efficacy of item parceling with different parceling strategies and different numbers of items per parcel, this systematic simulation study was conducted to examine the sensitivity of the fit measures to measurement noninvariance. The major manipulated design factors included parceling strategy (distributed or isolated), number of items per parcel, magnitude of measurement invariance, and sample size.

Method

The population model was a two-group two-factor CFA model with 24 items per factor (48 items in total), with Group 1 being the reference group and Group 2 being the focal group. All the indicators were continuous with a multivariate normal distribution. Factor loadings in the reference group and the metric invariant items in the focal group were set at 0.70 and residual variances were 0.51. Item intercepts in the reference group and the scalar invariant items were set at zero. These values were chosen based on the range of parameters used in previous research on measurement invariance testing using parcels (Meade & Kroustalis, 2006) and general measurement invariance testing using items (Cao & Liang, 2022; E. S. Kim et al., 2012; Liang & Luo, 2020). The correlation between the two factors was 0.30.

Design Factors

Location of Measurement Noninvariance. Measurement invariance was tested at two levels: factor loadings and item intercepts. These two locations of measurement invariance are the most frequently used in the literature of measurement invariance testing (Cao & Liang, 2022; Chen, 2007; E. S. Kim et al., 2012; Liang & Luo, 2020). To test the invariance of factor loadings, the model fit of configural and metric models was compared. Item intercepts invariance was tested by comparing the model fit of metric and scalar models.

Proportion of Items with Measurement Noninvariance. The proportion of noninvariant items was manipulated at 25% and 50% of the items for each of the two factors. These two proportions were consistent with those adopted by measurement invariance studies (Chen, 2007; Liang & Luo, 2020; Meade & Lautenschlager, 2004; Meade et al., 2008).

Magnitude of Measurement Noninvariance. The magnitude of measurement noninvariance was set at two different levels in this study. For both factor loadings and items intercepts, the magnitudes of noninvariant items in Group 2 were .15 and .30 lower than those in Group 1. That is, the factor loadings for metric noninvariant items in Group 2 were .55 and .40 for the two levels, respectively; the item intercepts for scalar noninvariant items in Group 2 were -.15 and -.30 for the two levels, respectively. The magnitude of measurement noninvariance was in the range employed in previous studies about measurement invariance (Cao & Liang, 2022; E. S. Kim et al., 2012; Liang & Luo, 2020; Meade & Kroustalis, 2006; Stark et al., 2006).

Sample Size. Two balanced groups with size of 200, 300, and 500 per group were examined in this study. The sample size reflected the range of sample size in empirical research about measurement invariance testing using individual items and item parcels (e.g., Glanville & Wildhagen, 2007; Meade & Kroustalis, 2006; Schlotz et al., 2011).

Number of Items Per Parcel. The number of items per parcel had five levels: 2, 3, 4, and 6 items per parcel so that 12, 8, 6, and 4 parcels per factor were obtained, respectively. These numbers of items per parcel were consistent with the literature that involves item parceling in the framework of SEM (Nasser & Wisenbaker, 2003). Also, individual items with no item parceling were also included in the analysis.

Parceling Strategies. Two parceling strategies were examined in this study: distributed and isolated parceling strategies. For isolated parceling, the noninvariant items were placed in the same parcels as much as possible. For example, when 25% of items (6 items per factor) were noninvariant, all the 6 items were put in the same parcel for 6-item parcel strategies. For 4-item parcel conditions, four noninvariant items were combined in the same parcel and the rest two noninvariant items were put in another parcel along with two invariant items, and the invariant items were put in the same parcels. For 3-item parcels, the 6 noninvariant items were put in two parcels. For 2-item parcels conditions, the 6 noninvariant items were put into three isolated parcels. The same principle was used in the conditions with 50% noninvariant items.

The opposite was true for distributed strategy: the noninvariant items were distributed to different parcels evenly as much as possible. For example, when 25% of items (6 items per factor) were noninvariant, for the 6-item parcels, 2 parcels contained two noninvariant items each, and the rest two parcels contained one noninvariant item each. For 4-item parcels conditions, the 6 parcels each contained one noninvariant item. For the 3-item parcels condition with 8 parcels, 6 parcels contained one noninvariant item each, and two parcels had no noninvariant items. For the 2-item parcels conditions with 12 parcels, 6 of the parcels contained one noninvariant item, and the rest 6 parcels contained no noninvariant item. The same principle was applied for conditions with 50% noninvariant items. Note that the 4-item parcels conditions with 6 parcels were the most evenly distributed conditions because the total noninvariant items were 6 or 12 per factor. Moreover, the parcels were correctly specified to load on their respective factors. Note that the models with individual items instead of item parcels were also analyzed to compare with models with different parceling strategies. The score of the parcel was created by taking the average of the items of the parcel.

In sum, there were 216 conditions (2 location of measurement noninvariance $\times$ 2 proportion $\times$ 2 magnitude $\times$ 3 sample size $\times$ (2 parceling strategy $\times$ 4 number of items per parcel + 1 no parceling) for noninvariant data generation. We also examined the false positive rates of detecting measurement noninvariance when the population data were invariant. The invariant data included two design factors: number of items per parcel, and sample size. Therefore, there were a total of 15 conditions of invariant data (5 number of items per parcel $\times$ 3 sample size).

Data Analysis

For each condition, 1,000 replications were generated and analyzed using Mplus 8.6 (Muthén & Muthén, 1998–2021). The item parcels were specified correctly to load on their respective factors. In configural invariance model, latent factor means and factor variances for both groups were fixed at zero and one, respectively, for model identification. Parcel intercepts, factor loadings, and residuals were freely estimated. In metric invariance model, the factor loadings of the parcels were constrained to be identical across the two groups. Factor means in both groups were fixed at zero. Factor variances in the reference Group 1 were fixed at one while freely estimated in Group 2. In scalar invariance model, parcel intercepts in addition to factor loadings were constrained to be equal across groups. Also, factor means and variances for Group 2 were freely estimated while being constrained to be zero and one in the reference group.

The proportion of replications in which the fit measures (CFI, RMSEA, LRT, AIC, BIC, and SaBIC) selected the configural invariance model over the metric invariance model or the metric invariance model over the scalar invariance model using the invariant data was computed as the false positive (FP) rate. For each replication of metric noninvariant data, configural and metric invariance models were analyzed. The proportion of replications that selected the configural invariance model over metric invariance model was computed as the sensitivity of the fit measures. For each replication of the scalar noninvariant data, metric invariance and scalar invariance models were analyzed. The proportion of replications that selected the metric invariance model over the scalar invariance model was considered as the sensitivity of the fit measures to scalar noninvariance. For model selection, evidence to support for measurement invariance included a non-significant LRT at the .05 alpha level, absolute CFI change < .01, RMSEA change < .015, and ICs change < 0 (more constrained model minus less constrained model).

Results

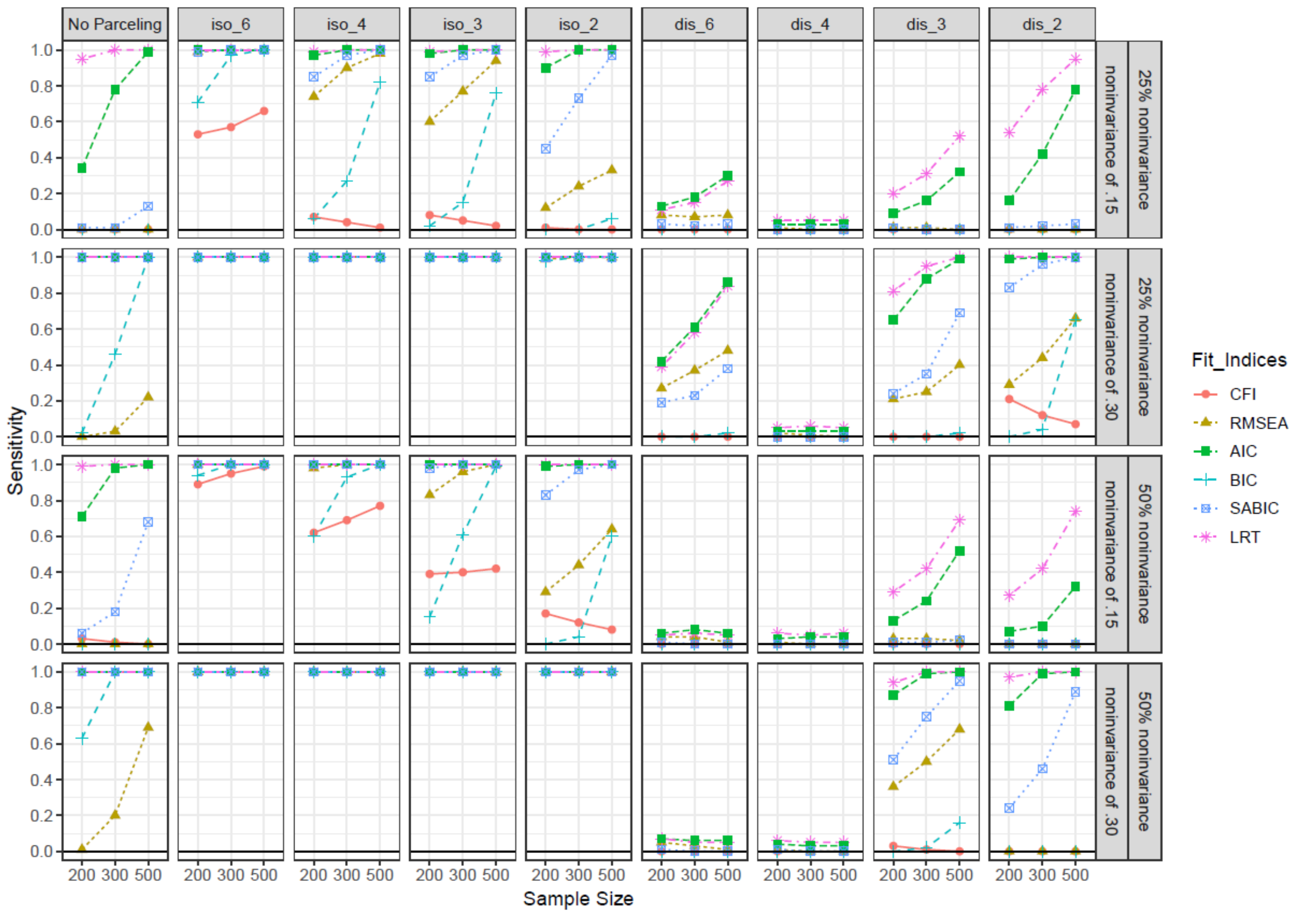
The false positive (FP) rates measurement invariance testing using the invariant data are summarized in Table 2. The FP rates for all the fit measures except LRT were mostly around zero. The FP rates of LRT were around .05. For AIC, the FP rates were mostly around zero except when there were 6 items per parcel, in which the FP rates were around .07. Given that the 95% confidence interval of the alpha level .05 based on 1,000 simulations is 3.7%-6.5% using the Clopper-Pearson method (Leemis & Trivedi, 1996), overall, the FP rates were under good control except for slight inflation for LRT and AIC in couple of conditions..

To determine the effect of the design factors (sample size, parceling strategy, number of items per parcel, magnitude of measurement noninvariance, and percentage of noninvariant items) on the sensitivity of each fit measures to measurement noninvariance, a factorial analysis of variance (ANOVA) with eta-squared (η^2) was conducted for the sensitivity of fit measures in metric invariance and scalar invariance testing. The percentage of variances explained by each design factor and their two-way interactions expressed as η^2 is presented in Table 3. Note that only the practically significant design factors and their interactions are reported based on the Cohen's (1988) moderate effect size of .0588, and any values of η^2 smaller than .0588 are not reported. Results indicated that parceling status explained a large portion of variance in RMSEA (around 79% and 81% for metric invariance and scalar invariance testing, respectively). CFI, BIC, and SaBIC in metric and scalar invariance testing were impacted by parceling status, the magnitude of noninvariance, and the interaction between parceling status and the magnitude of noninvariance. The sensitivity of LRT and AIC in both measurement invariance testing was impacted by parceling status, number of items per parcel, and the interaction between parceling status and the number of items per parcel.

Metric Invariance Testing

Figure 1 presents the sensitivity of all the fit measures in metric invariance testing. When no parceling was involved, CFI was not sensitive to the noninvariance of .15 but quite sensitive to the noninvariance of .30 regardless

Figure 1 ■ Sensitivity of Fit Measures to Metric Noninvariance with Different Parceling Strategies. In the figure, iso_2 = isolated parceling with 2 items per parcel; iso_3 = isolated parceling with 3 items per parcel; iso_4 = isolated parceling with 4 items per parcel; iso_6 = isolated parceling with 6 items per parcel; dis_2 = distributed parceling with 2 items per parcel; dis_3 = distributed parceling with 3 items per parcel; dis_4 = distributed parceling with 4 items per parcel; dis_6 = distributed parceling with 6 items per parcel.



of sample size. RMSEA was not sensitive to metric noninvariance in all conditions except in the condition with the largest sample size combined with 50% of noninvariance of .30. LRT was sensitive to metric noninvariance in all conditions. The sensitivity of AIC, BIC, and SaBIC depended on sample size and the magnitude of noninvariance.

In terms of the impact of different parceling strategies, for the magnitude of noninvariance of .30, isolated parceling increased the sensitivity of RMSEA and BIC to around 100% regardless of the number of items per parcel. For the magnitude of noninvariance of .15, isolated parceling was associated with higher sensitivity for all the fit measures compared to no parceling, especially for CFI, RMSEA, BIC, and SaBIC. Moreover, the increase of sensitivity of the fit measures was more drastic for more items per parcel. For example, the sensitivity of all the fit measures was much higher when there were 6 items per parcel than when there were 4 items per parcel, which was higher than when there were 3 and 2 items per parcel. The gradual increase in sensitivity of fit measures with the increase of number of items per parcel was evident for 50% of noninvariance of .15, particularly for CFI, RMSEA, and BIC.

On the other hand, distributed parceling seemed to decrease the sensitivity of all the fit measures. The decrease was more obvious with parcels with more items. LRT and AIC had the most dramatic decrease given their high sensitivity in conditions without parceling. When there were 2 or 3 items per parcel, the decrease in the sensitivity was not as much as in conditions with 4 and 6 items per parcel. It is worth noting that when there were 4 items per parcel, the

Table 2 ■ The False Positive Rates of No Parceling and Parceling with Different Number of Items Per Parcel

Sample	Parcel	CFI	RMSEA	LRT	AIC	BIC	SABIC
200	No_Parceling	.000	.000	.06	.000	.000	.000
200	item_2	.000	.000	.06	.000	.000	.000
200	item_3	.000	.000	.05	.012	.000	.000
200	item_4	.000	.012	.05	.034	.000	.002
200	item_6	.000	.048	.06	.076	.000	.018
300	No_Parceling	.000	.000	.07	.000	.000	.000
300	item_2	.000	.000	.05	.006	.000	.000
300	item_3	.000	.000	.04	.004	.000	.000
300	item_4	.000	.000	.05	.030	.000	.000
300	item_6	.000	.032	.05	.070	.000	.014
500	No_Parceling	.000	.000	.04	.000	.000	.000
500	item_2	.000	.000	.05	.000	.000	.000
500	item_3	.000	.000	.05	.010	.000	.000
500	item_4	.000	.000	.05	.022	.000	.000
500	item_6	.000	.012	.07	.076	.000	.002

Note. item_2 = parceling with 2 items per parcel; item_3 = parceling with 3 items per parcel; item_4 = parceling with 4 items per parcel; item_6 = parceling with 6 items per parcel.

Table 3 ■ Eta-squared (η^2) (%) of Different Sources for CFI, RMSEA, LRT, AIC, BIC, and SaBIC in Metric Invariance and Scalar Invariance Testing

	Metric Invariance Testing						Scalar Invariance Testing					
Source	CFI	RMSEA	LRT	AIC	BIC	SABIC	CFI	RMSEA	LRT	AIC	BIC	SaBIC
Parcel	51.79	79.09	53.46	57.55	60.54	67.23	48.8	80.74	54.21	58.91	61.66	70.55
Nonin	19.42			6.08	10.04	9.07	21.5				9.09	7.87
ItemPerParcel			15.07	8.37					14.76	7.94		
Parcel*Nonin	15.73				6.39	7.67	18.1				6.13	7.53
Parcel*ItemPerParcel			15.1	8.65					14.84	8.24		

Note. Parcel = parceling strategy; Nonin = the magnitude of noninvariance; ItemPerParcel = the number of items per parcel.

sensitivity of all fit measures was close to zero. This could stem from 4-item parcels featuring the most balanced distribution of noninvariant items (e.g., 1 noninvariant item per parcel across all 6 parcels in the 25% noninvariant item condition), whereas the distribution of noninvariant items varied more in other conditions. In sum, distributed parceling resulted in lower sensitivity of the all the fit measures, especially when there were more items per parcel.

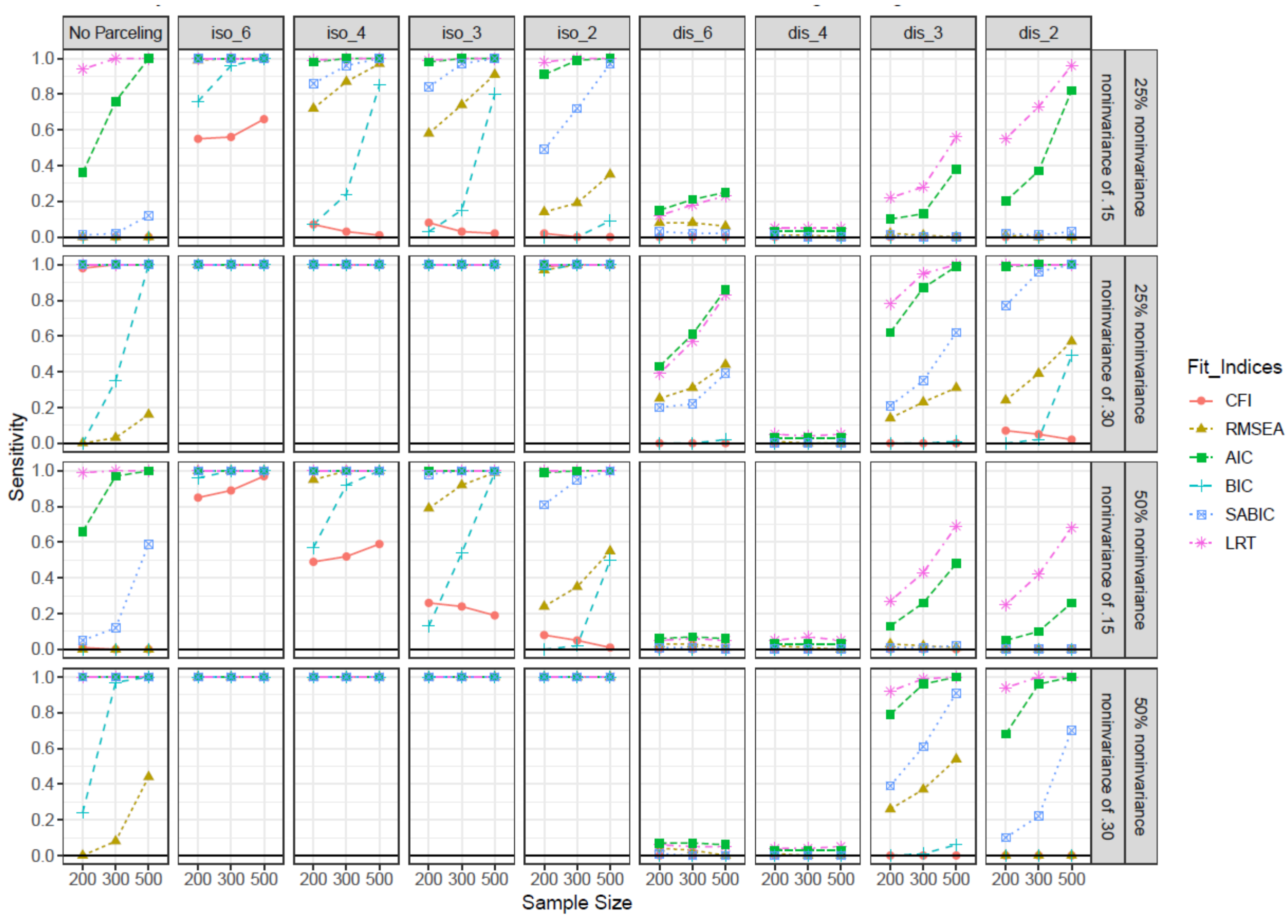
Scalar Invariance Testing

Figure 2 presents the sensitivity of all fit measures to scalar invariance testing. The pattern in scalar invariance testing was similar to the patten in metric invariance testing. The sensitivity of all fit measures increased with isolated parceling, especially when there were 6 items per parcel. The increase was particularly evident for fit measures that had low sensitivity without parceling, such as RMSEA, CFI, BIC, and SaBIC. The increase of sensitivity of all fit measures was less manifest with isolated parceling with fewer items per parcel, like 2 items per parcel. On the other hand, distributed parceling decreased the sensitivity of all the fit measures, especially when there were 6 or 4 items per parcel. For the most balanced 4-item parcels, the sensitivity of all fit measures in all conditions decreased to around zero. The decrease was less obvious in distributed parceling with 2 or 3 items per parcel.

Discussion

Item parceling has been used frequently in empirical research in SEM, including measurement invariance testing (e.g., Glanville & Wildhagen, 2007; Schlotz et al., 2011). However, there are limited discussions on the impact of item parceling on the performance of commonly used fit measures in measurement invariance testing. Thus, the study purported to contribute to this research gap by systematically examining the impact of parceling strategies and the number of items per parcel on the sensitivity of fit measures in measurement invariance testing through a

Figure 2 ■ Sensitivity of Fit Measures to Scalar Noninvariance with Different Parceling Strategies. In the figure, iso_2 = isolated parceling with 2 items per parcel; iso_3 = isolated parceling with 3 items per parcel; iso_4 = isolated parceling with 4 items per parcel; iso_6 = isolated parceling with 6 items per parcel; dis_2 = distributed parceling with 2 items per parcel; dis_3 = distributed parceling with 3 items per parcel; dis_4 = distributed parceling with 4 items per parcel; dis_6 = distributed parceling with 6 items per parcel.



simulation study. The findings from this study provided a more comprehensive picture of measurement invariance testing that involves parceling.

Not all parceling strategies "masked" the detection of measurement noninvariance, partially diverging from the findings of Meade and Kroustalis (2006), in which only the sensitivity of LRT was examined. When no parceling was used, the sensitivity of LRT was all around 100%, and its sensitivity remained around 100% in isolated parceling, but decreased substantially in distributed parceling. Therefore, it seemed that different parceling strategies yielded varying sensitivity to detecting measurement noninvariance for LRT. The impact of parceling depended on the parceling strategies and number of items per parcel. The results from this study provided applied researchers with a more comprehensive understanding of the impact of parceling in measurement invariance testing using parcels. Also, parceling has been recommended when structural relationships are of major interest to the researchers. When the structural relationships are examined with parcels, researchers may prioritize the invariance testing of parcels instead of individual items. Therefore, the potential masking of measurement noninvariance, which implies that the parcels exhibit measurement invariance, might be beneficial for estimating structural relationships since parcels are used as indicators of the latent factors instead of individual items.

Compared to no parceling, isolated parceling increased the sensitivity of all fit measures and the increase was more dramatic with more items per parcel, or fewer parcels. This was not unexpected given that isolated parceling combined all the noninvariant items into as few parcels as possible, enhancing the clarity of the parcels with respect to noninvariant and invariant items, making the noninvariance in the parcel(s) very apparent to detect. Also, isolated parceling can reduce measurement errors by averaging out errors across items within each parcel, potentially leading to more reliable parcel indicators. In addition, creating parcels reduces the model size. Previous research on the impact of model size on measurement invariance testing showed that the sensitivity of CFI, RMSEA, BIC, and SaBIC decreased when the number of observed indicators increased in the CFA model (Cao & Liang, 2022). This was consistent with the results that the sensitivity of fit measures increased as fewer parcels were created (i.e., decrease in model size) when the noninvariant items were parceled together.

On the other hand, distributed parceling decreased the sensitivity of all the fit measures, and the decreased sensitivity was more evident in parcels with 4 or 6 items per parcel. Particularly when there were 4 items per parcel and noninvariant items were evenly distributed across parcels, the sensitivity of all fit measures was close to zero in all conditions. The decreased sensitivity was less extreme in the parcels with 6 items per parcel, especially when there were 25% of noninvariant items. In general, when all the noninvariant items were evenly distributed to the parcels, the sensitivity of the fit measures diminished. Distributive parceling ensures that each parcel is representative of the entire spectrum of invariant and noninvariant item distribution. However, the inclusion of both invariant and noninvariant items within each parcel may reduce the parcels' interpretability.

Implications for Empirical Researchers

Based on the discussions above, for empirical researchers conducting measurement invariance testing, the strategic use of parceling can significantly impact the power to detect noninvariance. If preliminary analyses or theoretical considerations suggest that certain items behave differently across groups, researchers are recommended to consider isolated parceling to increase the sensitivity of detecting noninvariance by concentrating the noninvariant effects within specific parcels. Identifying noninvariant items to group together when using isolated parceling involves a multi-step process that combines statistical analyses with theoretical considerations. From a theoretical perspective, content experts may be able to provide good insight into the items' potential of resulting in measurement noninvariance based on the wording of the items, the context, and the target populations to collect data from. Also, literature and previous research on the same topic may also provide some hints about the possibility of having measurement noninvariance between groups. From a statistical point of view, preliminary analyses such as multigroup CFA with individual items, examination of modification indices, and regularized factorial invariance testing (Liang & Jacobucci, 2020; Shi et al., 2019), can help identify noninvariant items. Combining both the statical evidence and theoretical insights, decisions can be made on which items to form parcels by grouping identified non-invariant items together, aiming to maintain theoretical coherence within parcels. Then, one can conduct measurement invariance testing with the new parcel structure to validate the model. Additionally, conducting sensitivity analyses with various parceling strategies can illuminate the effects of parceling on measurement invariance detection, offering valuable insights into the robustness of the findings.

On the other hand, if the research focus is on improving the estimation of structural relationships, random parceling is likely to create evenly distributed parcels. These parcels, when used as indicators, are more likely to exhibit invariance especially with large sample sizes. In this simulation study, the parceling allocation was just one of many possible random allocations. Based on previous research, there has been variability in model fit indices across various allocations (Sterba & MacCallum, 2010; Sterba, 2011). That is to say, if we use a different item-to-parcel allocation, we may observe different model fit indices. However, Sterba and MacCallum (2010) showed that the variability of the model fit indices was most profound when sample size was small (below 200) and factor loadings were low. In this study, the sample size was relatively large, and the variability of model fit indices across different allocations was expected to be relatively small.

As for the fit measures to consider, LRT emerges as a preferred fit measure given its high sensitivity in measurement invariance testing using both individual items and parcels. When isolated parceling is used, AIC and SABIC may also be considered which perform similarly to LRT across most conditions. It should be kept in mind that the rationale for using item parcels instead of original items needs to be justified, and item parcels should not be used as a convenient alternative to original items. It is important to note that parceling, while useful, should not replace a substantive understanding of the construct and how it may be perceived or valued differently across groups.

Limitations and Future Research Directions

There are several limitations in this simulation study, which can serve as future research directions. First, the scenarios considered in this study were limited and the conditions can be expanded to include more simulation design factors. For example, the items simulated in this study are normally distributed continuous variables. However, in empirical research, it is likely to encounter nonnormal continuous variables or categorical variables. It remains unknown whether similar patterns of results are observed if nonnormal or categorical variables are used. Therefore, one of the future directions is to expand to categorical variables, and the impact of parceling strategies on measurement invariance testing using categorial data could be examined. Second, the sample-to-item ratio was 4.2 for the smallest sample size of 200 per group with 48 items. When sample size is smaller to be concerning, which may be one of the major reasons for using item parceling, the impact of parceling strategies on the sensitivity of fit measures in measurement invariance testing remains yet to know. In the future research direction, extremely smaller sample size can be used to investigate the convergence rates, and the impact of parceling strategies on measurement invariance testing and parameter estimates. Third, in the current study, the noninvariance was designed in the same direction (i.e., all the factor loadings or item intercepts in Group 2 were lower than Group 1). It is more likely to have some items with higher factor loadings or intercepts and some other items with lower loadings or intercepts than the comparison group. In this scenario, the noninvariance may be canceled off in a particular parcel that takes the average of several items. Thus, measurement invariance testing would be more likely to show that it holds. The scenarios designed in this study were the extreme example of noninvariance existing in the same direction. Lastly, this study focuses on the impact of parceling strategies on the sensitivity of fit measures in measurement invariance testing using CFA model. It is possible that the parceling strategies also influence the parameter estimates of structural relationships between latent factors in the context of SEM. Future research is called to examine the impact of parceling strategies on the estimates on structural parameters.

References

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, *19*, 716–723. doi: 10.1109/TAC.1974.1100705.

Bandalos, D. L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 78–102. doi: 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: A Multidisciplinary Journal*, *15*, 211–240. doi: 10.1080/10705510801922340.

Cao, C., & Liang, X. (2022). The impact of model size on the sensitivity of fit measures in measurement invariance testing. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*, 744–754. doi: 10.1080/10705511.2022.2056893.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*, 464–504. doi: 10.1080/10705510701301834.

Cheung, G. W., & Rensvold, R. B. (2001). The effects of model parsimony and sampling error on the fit of structural equation models. *Organizational Research Methods: A multidisciplinary Journal*, *4*, 236–264. doi: 10.1177/109442810143004.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum.

Glanville, J. L., & Wildhagen, T. (2007). The measurement of school engagement: Assessing dimensionality and measurement invariance across race and ethnicity. *Educational and Psychological Measurement*, *67*, 1019–1041. doi: 10.1177/0013164406299126.

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

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

Kenny, D. A., & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*, 333–351. doi: 10.1207/S15328007SEM1003_1.

Kim, E. S., Yoon, M., & Lee, T. (2012). Testing measurement invariance using mimic: Likelihood ratio test with a critical value adjustment. *Educational and Psychological Measurement*, *72*, 469–492. doi: 10.1177/0013164411427395.

Kim, S.-H., Cohen, A. S., Cho, S.-J., & Eom, H. J. (2019). Use of information criteria in the study of group differences in trace lines. *Applied Psychological Measurement*, *43*, 95–112. doi: 10.1177/0146621618772292.

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. doi: 10.1177/109442810032003.

Lee, J., & Whittaker, T. (2021). The impact of item parceling on structural parameters invariance in multigroup structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*, 684–698. doi: 10.1080/10705511.2021.1890604.

Leemis, L. M., & Trivedi, K. S. (1996). A comparison of approximate interval estimators for the bernoulli parameter. *The American Statistician*, *50*(1), 63–68. doi: 10.1080/00031305.1996.10473544.

Liang, X., & Jacobucci, R. (2020). Regularized structural equation modeling to detect measurement bias: Evaluation of lasso, adaptive lasso, and elastic net. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*, 722–734. doi: 10.1080/10705511.2019.1693273.

Liang, X., & Luo, Y. (2020). A comprehensive comparison of model selection models for testing factorial invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*, 380–395. doi: 10.1080/10705511.2019.1649983.

Little, T. D. (1997). Mean and covariance structures (macs) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research*, *32*, 53–76. doi: 10.1207/s15327906mbr3201_3.

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: A Multidisciplinary Journal*, *9*, 151–173. doi: 10.1207/S15328007SEM0902_1.

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

Marsh, H. W., Hau, K. T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? the number of indictors per factors in confirmatory factor analysis. *Multivariate Behavioral Research*, *33*, 181–222. doi: 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. doi: 10.1037/a0032773.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*, 568–592. doi: 10.1037/0021-9010.93.3.568.

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. doi: 10.1177/1094428105283384.

Meade, A. W., & Lautenschlager, G. J. (2004). A monte-carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*, 60–72. doi: 10.1207/S15328007SEM1101_5.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543. doi: 10.1007/BF02294825.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.

Muthén, L. K., & Muthén, B. O. (1998–2021). *Mplus user's guild* (8th ed.). Muthén & Muthén.

Nasser, F., & Wisenbaker, J. (2003). A monte carlo study investigating the impact of item parceling on measures of fit in confirmatory factor analysis. *Educational and Psychological Measurement*, *63*, 729–757. doi: 10.1207/s15328007sem1302_3.

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. doi: 10.1207/S15327906MBR3903_1.

Schlotz, W., Yim, I. S., Zoccola, P. M., Jansen, L., & Schulz, P. (2011). The perceived stress reactivity scale: Measurement invariance, stability, and validity in three countries. *Psychol Assess*, *23*, 80–94. doi: 10.1037/a0021148.

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

Sclove, S. L. (1987). *Application of model-selection criterion to some problems in multivariate*.

Shi, D., Song, H., DiStefano, C., Maydeu-Olivares, A., McDaniel, H. L., & Jiang, Z. (2019). Evaluating factorial invariance: An interval estimation approach using Bayesian structural equation modeling. *Multivariate Behavioral Research*, *54*(2), 224–245. doi: 10.1080/00273171.2018.1514484.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, *91*, 1292–1306. doi: 10.1037/0021-9010.91.6.1292.

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

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

Vrieze, S. I. (2012). Model selection and psychological theory:a discussion of the differences between the akaike information criterion(aic) and the Bayesian information criterion (bic). *Psychological Methods*, *17*, 228–243. doi: 10.1037/a0027127.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain [Psychometrika, 52]. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 333–343). American Psychological Associationanalysis.

Citation

Cao, C., & Liang, X. (2025). Exploring the effect of item parceling strategies and number of items per parcel on measurement invariance testing in confirmatory factor analysis. *The Quantitative Methods for Psychology*, *21*(2), 95–107. doi: 10.20982/tqmp.21.2.p095.

Received: 13/03/2025 ∼ Accepted: 28/06/2025