Cross-Loadings in Scale Development: Monte Carlo Study of Structural Item-Total Correlation Analysis with Small Samples

Gordon P. Brooks
Ohio University
James Sika Pokoo
Ohio University
Nina Adjanin
Northwest Missouri State University
George A. Johanson
Ohio University

This study investigated the possibility of using item correlations with subscales as a tool to diagnose cross-loadings in the scale development process using smaller sample sizes than required for factor analysis. A Monte Carlo simulation using R examined sample sizes from 30-120 under several conditions of correlations, numbers of items (8-36), and numbers of factors (2-3). Within each condition, most items were generated to load on their expected dimensions, but some items were generated that (a) cross-loaded on multiple dimensions or (b) loaded on no dimensions. Based on its consistently larger average loading differences between loading power (loading correctly on the right dimension) and loading error (wrongly loading on an incorrect dimension), combined especially with its consistently lower loading errors, Structural Item-Total Correlation Analysis (SITCA) diagnosed cross-loading and non-loading items most effectively across most of the conditions when sample sizes were approximately 50-60.

The importance of quality items in scale development cannot be overestimated. However, in the current world, samples are increasingly difficult to obtain, making the process of scale development more difficult. Potentially, a single new scale could conceivably require multiple samples for pilot studies and validity studies to improve items and provide evidence of validity and reliability, respectively, depending on how many changes are made to the scale during the process (ideally, item analyses and validity studies continue after any items are changed before the scale is used for applied research purposes). Having the ability to weed out and improve items that need to be repaired or replaced (or removed) as early as possible in the process, with typically smaller pilot study samples, allows for fewer more extensive and more expensive validity study samples later.

Applied researchers perform a number of statistical analyses as they develop tests and scales, in order to provide evidence of both reliability and validity. For example, they perform item analyses (e.g., alpha, item-total correlations, alpha-if-item-deleted) for unidimensional scales and subscales to verify that all items are contributing positively to reliability. Previous authors have suggested that preliminary or pilot studies for these issues in scale development can be performed successfully with small sample sizes of 24-36 cases based on the precision of correlations (standard errors, confidence intervals) used in item analyses (Johanson & Brooks, 2010). Evidence (often from experts) for content validity arguments is also needed.
Scale developers also typically perform factor analyses for structural validity to provide evidence that the underlying dimensional structure of an instrument supports construct validity. These factor analytic methods—for example, exploratory factor analysis (EFA), confirmatory factor analysis (CFA), or principal components analysis (PCA)—are also the best methods to verify (a) that items are contributing to the measurement of their designated subscale and (b) that items are not contributing to multiple subscales (item unidimensionality). Items that contribute significantly to more than one subscale (or factor or component) are commonly called cross-loaded items. Therefore, in the scale development process, while still creating and revising items, researchers often desire to use factor analysis to help examine the quality of items: whether items load statistically on any of the dimensions of the scale, whether items load on the correct dimensions, and whether items load on just one dimension. This evidence is useful as the scale and items are still being developed so the items can be improved as much as possible before the more extensive construct validity evidence is collected.

Unfortunately, these factor analytic methods generally require hundreds of cases for most scales of decent size and with multiple subscales. For example, some say minimum sample sizes of at least 150 should be used even in the very best circumstances (Bandalos & Finney, 2010; Gorsuch, 1983; Guadagnoli & Velicer, 1988; Kahn, 2006; Kline, 1994). Frequently, however, scholars simply suggest samples as large as possible, perhaps over 300 cases or cases-to-items ratios such as 10:1 or 20:1, which would require 200 or 400 cases, respectively, for just 20 items (Comrey & Lee, 1992; Hair et al., 1998; Kline, 2016; Tabachnick & Fidell, 2007). Mundfrom, Shaw, and Ke (2005) and Pearson and Mundfrom (2010) reported

that with larger item-to-factor ratios like 7:1 to 8:1, minimum sample sizes required for most circumstances were generally below 150 when there were high communalities. They noted, however, that many more cases may be required with smaller item-to-factor ratios. Ultimately, these authors suggested that in the best of conditions (at least seven items per factor and high communality), as few as 60 cases may be required for factor analysis. De Winter et al. (2009) provided some guidelines for EFA with small samples, also emphasizing the importance of high communalities and large item-to-factor ratios.

Purpose of the Study

The purpose of this study was to investigate the possibility of using item-total correlations as a proxy for factor analysis with much smaller sample sizes that might be more easily obtainable during pilot study work early in the scale development process. We present a method we call "Structural Item-Total Correlation Analysis" (SITCA) that may help researchers use smaller samples earlier in the process to identify items that may be problematic in ways related to the structure of the scale. SITCA describes an item-total correlations analysis process analogous to factor analysis, not to replace EFA, CFA, or PCA, but rather as a method to analyze relationships between items and composite scores on subscales—with the specific potential to identify obviously troublesome items and to do this with smaller sample sizes.

Note that for both theoretical and philosophical reasons, we—like many others—prefer EFA and CFA to PCA for structural validity and scale and item analysis, but EFA and CFA often face difficulties with small samples. Therefore, for practical reasons, we used PCA for comparative purposes in this study. Thankfully, studies have suggested that PCA and EFA provide equivalent results when underlying dimensional structures are relatively strong, as often can be expected in scale development (Tabachnick & Fidell, 2007; Velicer & Jackson, 1990). Further, scholars often performed exploratory EFA or PCA and gave it a confirmatory interpretation before CFA became more commonly used. However, even now, we might argue that CFA is often used prematurely, before the quality of items has been thoroughly studied—that researchers should ensure item quality before moving to CFA, with larger more costly sample sizes, thereby ensuring that items are strong before testing theory.

Factor Analysis in Scale Development

In PCA and EFA, researchers assume that items within subscales are relatively strongly correlated—and more highly correlated with items from their own subscales than with items from other subscales. PCA and EFA are exploratory in that they use data to show where structures or dimensions are located; CFA is confirmatory because the structure is assumed known and the question is whether the data fit this known structure. Usually, the number of factors to be extracted in EFA and PCA is based on parallel analysis, minimum average partial (MAP), Kaiser's rule, or scree plots, among other methods (Tabachnick & Fidell, 2007). However, in scale development, researchers create instruments and items designed to measure particular aspects of a larger construct, that is, designated to measure a particular subscale—for example, using theory, a table of specifications, or scale blueprint. Therefore, because the researcher has a keen sense of how many dimensions there should be, that known number may be used to extract factors when performing EFA or PCA—and is required for defining the measurement model for CFA. That is, researchers extract the number of components equal to the number of subscales intended or designed in the scale. The exploratory or confirmatory factor analytic results then help to verify that each item is associated statistically (loads) with (or "belongs to") the dimension of the construct (factor, component, or subscale) it is intended to be associated with—its designated dimension.

Researchers use loadings, which are correlations (or standardized regression weights or path coefficients), observed between the item variables and latent common factors or composite components to examine the strength of these associations (Bai, & Ng, 2008; Hair et al., 1998). Rotated loadings are often used to help the set of items load more strongly with the extracted factors, with oblique rotation most commonly used in scale development due to the expected correlations among the dimensions of a construct (Bandalos, 2018; Bandalos & Finney, 2010; Hair et al., 1998; Pituch & Stevens, 2015; Rummel, 1970; Tabachnick & Fidell, 2007). In oblique rotation (e.g., Promax or Direct Oblimin) there are two matrices of loadings: the pattern matrix of regression weights that represent the unique relationship of a factor to an item with all other factors held constant (recommended by Fabrigar & Wegener, 2012) and the structure matrix of correlations between items and factors that assess both those unique relationships and the additional association introduced by any overlap among the factors.

In Figure 1 we present familiar loadings tables from PCA for a two-dimensional scale: unrotated component matrix (Figure 1a), obliquely rotated pattern matrix (Figure 1b), and obliquely rotated structure matrix (Figure 1c). In the Pattern Matrix in Figure 1b, we see that most items load as expected from this generated data (N = 120): V1-V5 load on Component 2 and V7-V11 on Component 1. These 10 items exhibit "simple structure": high loadings above some criterion on only one component, low loadings on other components, and components having sufficient items for strong reliability. When factor loadings show that all items each load highly on only one dimension, this is called simple structure (Bandalos & Finney, 2010; Thurstone, 1947), which is often the ideal pattern of loadings desired.

Various rules for these "high" loadings have been considered by scholars. For example, Pituch and Stevens (2015) suggested a threshold of greater than .40, where the greater the loadings, the more the item or variable is a pure measure of the factor (Bandalos, 2018; Bandalos & Finney, 2010; Comrey & Lee, 1992; Hair et al., 1998; Pituch & Stevens, 2015; Tabachnick & Fidell, 2007). Some of these scholars have categorized based on the squared loadings as measures of overlapping variance between the factors and the items, with loadings of .32 (10% shared variance) as the minimum for an acceptable loading. Jordan and Spiess (2019) reported that the second standardized loadings should be below a cutoff of .30. However, most authors do not discuss the maximum size of the second highest loadings (necessary for diagnosing cross-loading violations to simple structure) in these conventional rules, but sometimes define or imply cross-loading as two loadings above these thresholds. However, adherence to such threshold rules can result in arbitrarily small differences between loading and not-loading, for example, .41 considered loading but .39 not.

Returning to Figure 1b, two items (V6 and V12) do not load clearly on just one component because both items' rotated pattern coefficients are relatively high for both components (comparable results appear in the rotated structure matrix, Figure 1c). That is, V6 and V12 are somewhat equivalently associated with both components. Consequently, we have concerns about the quality and unidimensionality of V6 and V12 and would not yet want to undertake validity studies before we can repair, replace, or (as a last resort) remove these items. Likewise, if any items have small loadings on both components, those items would not be considered to measure either component ("not-loaded"). No examples of not-loading items are shown, but the reader can easily imagine a hypothetical V13 with a .1 loading on both components.

(a)

Component Matrix[a]

	Component	
	1	2
V1	.522	.483
V2	.487	.501
V3	.574	.433
V4	.553	.406
V5	.503	.497
V6	.761	
V7	.703	-.320
V8	.652	-.322
V9	.603	-.481
V10	.646	-.494
V11	.687	-.278
V12	.837	

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

(b)

Pattern Matrix[a]

	Component	
	1	2
V1		.740
V2	-.110	.742
V3		.714
V4		.677
V5		.745
V6	.518	.371
V7	.766	
V8	.733	
V9	.840	-.196
V10	.880	-.189
V11	.717	
V12	.561	.418

Extraction Method: Principal Component Analysis. Rotation Method: Promax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

(c)

Structure Matrix

	Component	
	1	2
V1	.264	.708
V2	.224	.692
V3	.332	.719
V4	.325	.686
V5	.240	.702
V6	.685	.604
V7	.773	.360
V8	.727	.318
V9	.751	.182
V10	.795	.208
V11	.739	.373
V12	.749	.671

Extraction Method: Principal Component Analysis. Rotation Method: Promax with Kaiser Normalization.

Figure 1. Example Loadings from PCA with N = 120

SITCA: Structural Item-Total Correlation Analysis

We present SITCA as a proxy for factor analysis to help identify cross-loaded or not-loaded items. Because SITCA is based on imposing fundamentally confirmatory structural assumptions to data analyzed with simple item-total correlations, but without the burden of estimating latent variables, sample size requirements will not be as large as for confirmatory factor analysis. The PCA structure matrix provides item-component correlations (linear combinations of all items, regardless of which component they belong to), but SITCA provides item-total correlations between items and a summated or averaged score for their own designated scale only (in this way, a bit more like CFA than EFA or PCA). Item-total correlations are often referred to item discrimination indices and the relationship of discrimination indices to factor loading is well known (Richardson, 1936; Henrysson, 1962; Jordan & Spiess, 2019).

In Figure 1, items V1-V6 are expected a priori (based on the item generation process) to be part of Component 2 and items V7-V12 are expected to be Component 1. Figure 2 shows these item-total correlations (not corrected or adjusted item-total correlations). For example, Scale 1 was added to the data for this sample as the average scores for only items V1-V6—not including items V7-V12, so we see higher correlations between items V1-V6 and Scale1. Similarly, items V7-V12 are highly correlated with Scale2, which was calculated as the average of only items V7-V12. We have arbitrarily numbered the scales whereas PCA extracts components in order based on variance, therefore Scale1 is the same as Component 2 in our example. We also see high correlations for items V6 and V12 with both scales.

Correlations[b]

		scale1	scale2
V1	Pearson Correlation	.675**	.286**
	Sig. (2-tailed)	.000	.002
V2	Pearson Correlation	.679**	.241**
	Sig. (2-tailed)	.000	.008
V3	Pearson Correlation	.710**	.338**
	Sig. (2-tailed)	.000	.000
V4	Pearson Correlation	.719**	.312**
	Sig. (2-tailed)	.000	.001
V5	Pearson Correlation	.695**	.248**
	Sig. (2-tailed)	.000	.006
V6	Pearson Correlation	.685**	.611**
	Sig. (2-tailed)	.000	.000
V7	Pearson Correlation	.381**	.775**
	Sig. (2-tailed)	.000	.000
V8	Pearson Correlation	.326**	.738**
	Sig. (2-tailed)	.000	.000
V9	Pearson Correlation	.255**	.723**
	Sig. (2-tailed)	.005	.000
V10	Pearson Correlation	.267**	.784**
	Sig. (2-tailed)	.003	.000
V11	Pearson Correlation	.367**	.765**
	Sig. (2-tailed)	.000	.000
V12	Pearson Correlation	.628**	.785**
	Sig. (2-tailed)	.000	.000

Figure 2. Example Item-Total Correlations with N = 120 (not corrected or adjusted by deleting each item from its scale score).

Methods and Data Sources

This study used Monte Carlo methods in R to generate and analyze data that fit certain conditions. We performed 10,000 simulated samples for each condition. We generated samples ranging from 30-120 and conditions where (a) some items load on no components (not-loaded), (b) some items cross-load on multiple components (cross-loaded), and (c) most items load as expected. The study included simulations for different numbers of items (8, 12, 16, 18, 20, 24, 30, 36) and different numbers of dimensions (2, 3). Population correlation matrices were defined to represent relatively strong simple structures and reflected varying sizes of correlations between items and their designated components (correlations of items with other items on the same subscale of, for example, .4 or .5) versus between items and other components (correlations of items with items on different subscales of, for example, .2 or .3).

Multivariate normal datasets were generated from these correlation matrices as population data with 2,000,000 cases for each condition and then random samples were drawn from the datasets. The data mimicked ordinal integer scale data from 1 to 5. We generated data based on correlation matrices where the sizes of the correlations differed by 0.1, 0.2, and 0.3. To carry the examples above further, we generated some correlation matrices where the correlations among items within their designated subscales were .3, .4, and .5 while the correlations with items not in the same subscales were .2 (for example, see Table 1, which includes Promax obliquely rotated loadings related to that correlation matrix).

The SITCA and PCA approaches necessarily work in conjunction with numeric rules that are used to determine minimum and maximum acceptable loadings to determine simple structure. We adapted simple structure criteria from our experience to determine how well SITCA and PCA correctly identified not-loaded/cross-loaded items. However, we did not find many such rules for cross-loading in the literature, as most rules were like those mentioned earlier for single loadings rather than for combinations of loadings for items. Therefore, we examined the quality of many rules and report our results about SITCA and PCA approaches based on the rules that worked most effectively across the most conditions.

The rules we ultimately used were based on the magnitudes of the correlations or loadings across the dimensions. These rules were among the most useful for the approaches we investigated: SITCA, PCA

Table 1. An example population correlation matrix for data generation with correlations r = .4 among items on the same subscale (shaded yellow and green) and correlations with items on another subscale r = .2 (a correlation differential, r_diff, = 0.2), with loadings from PCA with Promax rotation.

	Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8	Loading1	Loading2
Item1	1.0	0.4	0.4	0.4	0.2	0.2	0.2	0.2	.790	.001
Item2	0.4	1.0	0.4	0.4	0.2	0.2	0.2	0.2	.790	.001
Item3	0.4	0.4	1.0	0.4	0.2	0.2	0.2	0.2	.790	.001
Item4	0.4	0.4	0.4	1.0	0.2	0.2	0.2	0.2	.790	.001
Item5	0.2	0.2	0.2	0.2	1.0	0.4	0.4	0.4	.001	.790
Item6	0.2	0.2	0.2	0.2	0.4	1.0	0.4	0.4	.001	.790
Item7	0.2	0.2	0.2	0.2	0.4	0.4	1.0	0.4	.001	.790
Item8	0.2	0.2	0.2	0.2	0.4	0.4	0.4	1.0	.001	.790

Note*: Sample correlation matrices would not be nearly so strong in simple structure as the population, particularly with smaller sample sizes.*

pattern loadings, and PCA structure loadings. The first sets of results we present below focus on the six rules that appeared to work best across the most conditions: (a) items load with a coefficient or correlation of at least .40, .50, or .60 on their designated subscales and this larger coefficient or correlation is at least 0.2 greater than loadings on other subscales (called .4-.2, .5-.2, and .6-.2, respectively); and (b) items load with a coefficient or correlation of at least .40, .50, or .60 on their designated subscales and this larger correlation is at least 1.25 times greater than loadings on other subscales (called .4>1.25x, .5>1.25x, and .6>1.25x, respectively). In these results, we also included the rule that items load with a coefficient or correlation of at least .60 on their designated subscales and no other loadings with other dimensions are above .40 (called .6v.4). In later results we focus just on the .5-.2 and .5>1.25x rules because they performed most effectively in the most conditions. We tried other rules that did not perform as well (and therefore we do not report them because these rules were not the focus of this research), for example, rules like .5-.3, .5>2x, .5>1.5x, and rules based on significance or communalities, among others.

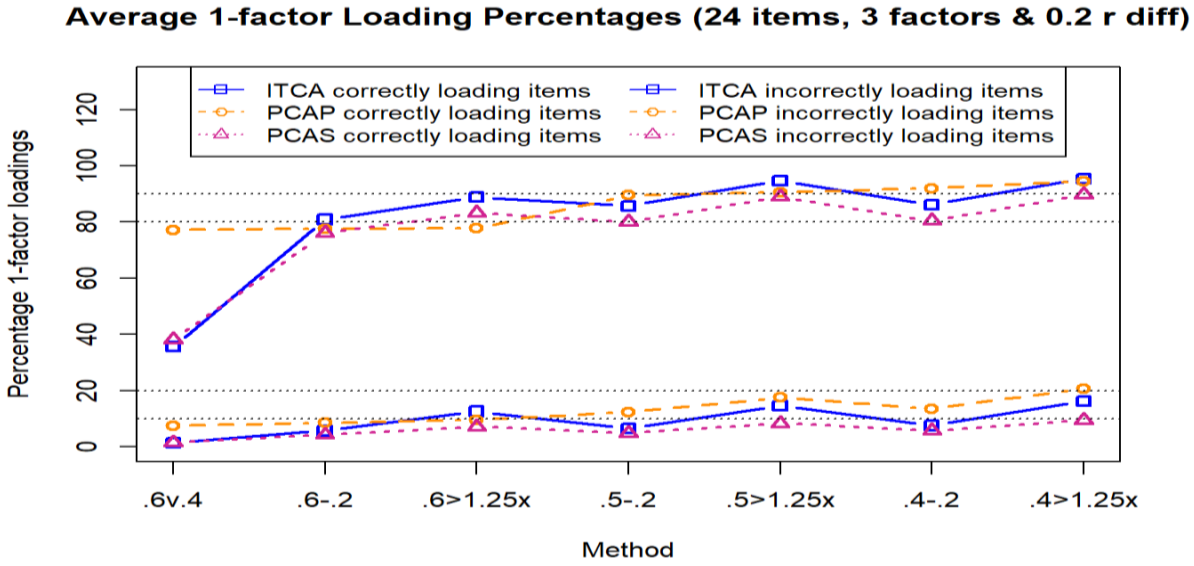
In each condition, we counted the number of samples in which items loaded cleanly on one dimension. Consequently, we looked for high percentages of samples where items correctly loaded on the one dimension they were designated to load on (these are "correctly loading" results, like power, that we will call "loading power"—and conversely, where an item failed to load where it should load would be akin to a Type II error). Similarly, we looked for low percentages of samples where items incorrectly loaded on a single dimension when they were designed to load not-at-all or designed to cross-load (these are "incorrectly loading" results, like Type I errors, that we call "loading errors" because they should not have loaded on just one dimension).

Results and Conclusions

We present results that were similar and relatively consistent across conditions. Multiple correlation matrices and conditions were used to generate data, so where appropriate, these results represent averages across conditions (we verified that means reported in this way provided a reasonable sense of the results). Unsurprisingly, most results got better as sample sizes increased. We found that the 0.1 correlation differentials provided too little structure and made both SITCA and PCA analyses more ambiguous—errors at both ends were generally too large to be acceptable with smallest sample sizes (but the methods worked effectively with larger sample sizes). However, our experience suggests that 0.2 or 0.3 or larger correlation differences between items and their designated dimensions versus other dimensions (r_diff = 0.2 and r_diff = 0.3 in the figures) is not unreasonable in scale development, especially after pilot or preliminary studies have sharpened the items used in the subscales.

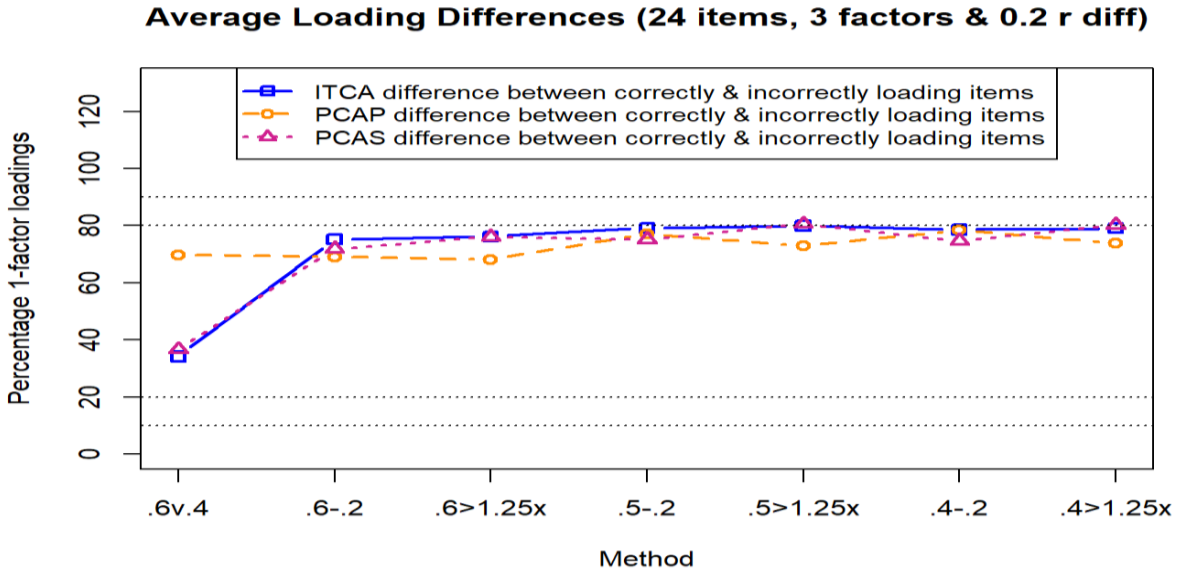
Figure 3 is one example condition, averaged across sample sizes, that shows both the (upper) percentages of samples where the methods showed correlations or loadings for items on the correct designated dimensions as well as the (lower) percentages for which the methods showed where items load on a single dimension but should not have. So, for example, with the .6-.2, .6>1.25x, .5>1.25x, and .4>1.25x rules in this Figure 3 condition, SITCA had the highest loading power. Both SITCA and PCAS had almost

Figure 3. Loading Power (higher) percentages and Loading Error (lower) percentages averaged across sample sizes for the condition with 24 items on 3 factors with 0.2 differences between correlations for designated factors versus other factors.



Note*: SITCA represents Structural Item-Total Correlation Analysis, PCAP represents PCA Loadings analysis based on Pattern Matrix, PCAS represents PCA Loadings analysis based on Structure Matrix.*

Figure 4. Differences in Loading Power percentages and Loading Error percentages averaged across sample sizes for the condition with 24 items on 3 factors with 0.2 differences between correlations for designated factors versus other factors



Note: SITCA is Structural Item-Total Correlation Analysis, PCAP is PCA Loadings analysis based on Pattern Matrix, PCAS is PCA Loadings analysis based on Structure Matrix

no loading errors with the .6v.4 rule, but both had unacceptably low loading power with this rule. We observed this general pattern or trade-off across most conditions—that is, the methods with higher loading power also tended to have higher loading error for that condition (and lower power with lower error – again, not unlike power and Type I error for many statistics, where lower alpha means lower power). There is clearly a trade-off in the choice of methods for maximizing loading power versus minimizing loading error.

Because we desire methods with higher loading power percentages (perhaps over 80%) and also lower loading error percentages (perhaps under 20%), we provide Figure 4 that shows average differences between these two percentages. Using Figure 3, we can see that SITCA has roughly 95% loading power

with the .5>1.25x rule and approximately 15% loading error with that same rule in this condition. In Figure 4, we see this average difference of approximately 80% (i.e., 95%-15%). We can also see that despite having lower loading power, because it also has lower loading error, PCAS has essentially the same average loading difference as SITCA for this condition. Larger average differences showed more deviation (which is good) between higher loading power and lower loading error results.

Figures 5a and 5b show results for the six primary methods (two simple structure rules used with three types of coefficients) for two sample sizes under one condition, with Figure 5c showing results averaged across all sample sizes for that condition. As mentioned above, the .5-.2 and .5>1.25x rules tended to provide the best performance across the most conditions. These figures show the data that serves as the input for the next series of graphs (Figures 6-14). Note that Figure 5c also shows the minimum differences for each rule and method combination. These represent the smallest (worst) difference between the loading power and loading error across all the sample sizes for that condition. Like average differences, larger minimum differences are also better. In our review of the results, we found essentially the same conclusions for comparing methods whether using the average loading differences or the minimum differences.

We provide Figures 6-14 to show results across several conditions and sample sizes. Figures 6-8 show the average loading differences (power versus error) for 4:1, 6:1, and 8:1 item-to-factor ratio conditions across the sample sizes from 30-120, respectively. Figures 9-11 and 12-14 show similarly the minimum loading power across the conditions and the maximum loading errors, respectively. For all three results (average loading differences, loading power, and loading errors), we found that above an 8:1 item-to-factor ratio the changes are relatively minimal and therefore do not show above 8:1 (patterns and rankings of more effective methods remain essentially the same).

By comparing the difference in correlations (i.e., r_diff) values of 0.2 and 0.3 across Figures 6-14, we can see clearly that having larger correlations among subscale items as compared to correlations across items of different scales results in higher loading power, lower loading error, and larger average differences between the two. Because of this, smaller sample sizes are needed to achieve any given level of loading power or minimized loading error with larger correlation differentials. For example, comparing Figures 8a and 8b, we can see that all rules hit the 80% average difference at roughly $N = 70$ for a difference of 0.2 but at approximately $N = 40$ for a difference of 0.3 among correlations. We see a similar pattern when we examine loading power (Figures 11a and 11b) and loading error (Figures 14a and 14b) separately. Figures 6-8 show that across most of the conditions, the SITCA with the .5-.2 simple structure rule was among the largest, if not the largest, average difference between loading power and loading error rates.

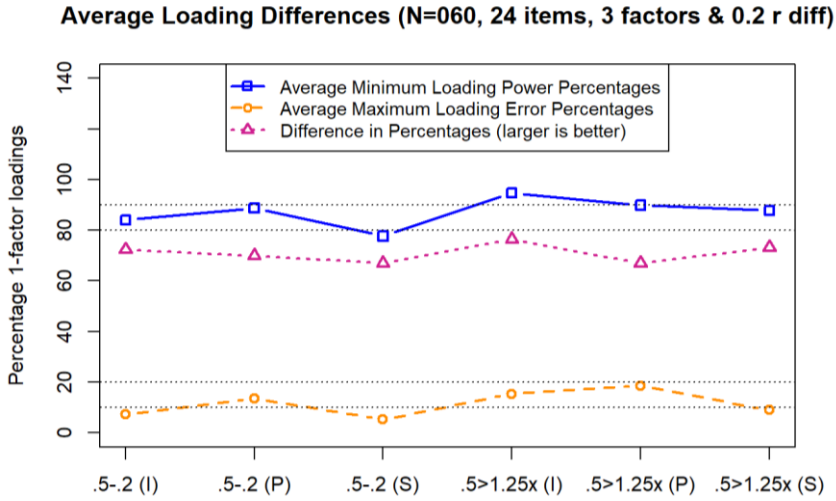
Looking across Figures 9-11, it is also relatively clear that small sample sizes may be sufficient if using 80% as the criterion for acceptable minimum loading power in many conditions. Indeed, in most of the conditions when the coefficient or correlation differential is 0.3 (r_diff = 0.3), sample sizes as small as 30-40 may be sufficient for most conditions using any of the methods. When the correlation differential is 0.2 (r_diff = 0.2), however, sample sizes of 50-70 are required for all methods to reach the 80% criterion. If loading power is the primary interest, then PCAS with the .5>1.25x rule shows the highest loading power across most conditions, with PCAP and .5>1.25x having the second highest loading power. Curiously, loading power does not tend to improve much as the item-to-factor ratios increase from 4:1 to 8:1 (recall that above 8:1 the results were essentially the same, with the methods becoming more tightly similar graphically).

In Figures 12-14, larger samples are clearly required to keep maximum loading errors low for most methods. Indeed, some of the approaches never result in smaller than 20% loading errors, even as high as $N = 120$. The 4:1 item-to-factor ratio results were particularly unhappy in this regard, suggesting that larger samples are required when subscales have fewer items, even for the best method (SITCA required at least 60-70 cases to get as low as 20% loading errors in the 4:1 ratio conditions with r_diff = 0.3, and did not reach 20% even with 120 cases when r_diff = 0.2). However, loading errors tended to decrease as item-to-factor ratios increased (such as comparing Figures 12c, 13c, 14c). SITCA with the .5-.2 rule showed the fewest loading errors across all conditions and sample sizes, often 5% lower than the next best method. With its lower loading errors, SITCA with the .5-.2 rule also consistently provided among the largest (if not the largest) average loading differences among the methods tested (see Figures 6-8).

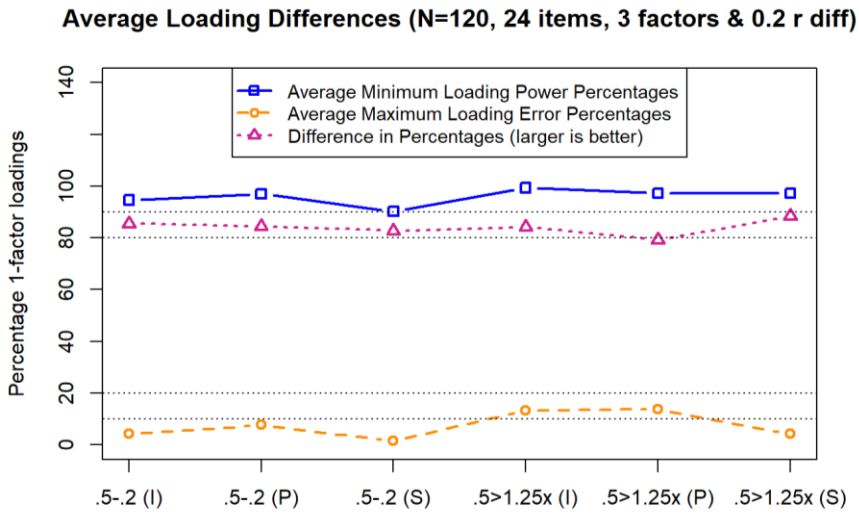
Lower loading errors with reasonable loading power may be the most important key to identifying not-loaded/cross-loaded items. That is, with lower loading errors, not-loaded/cross-loaded items are less likely to load on a single dimension, and therefore easily diagnosed. We do not want items that do not belong to

Figure 5. Average Differences in Loading Power (higher) percentages and Loading Error (lower) percentages at (a) N = 60, (b) N = 120, and (c) across all N conditions for conditions with 24 items on 3 factors with 0.2 differences between correlations within designated factors versus other factors.

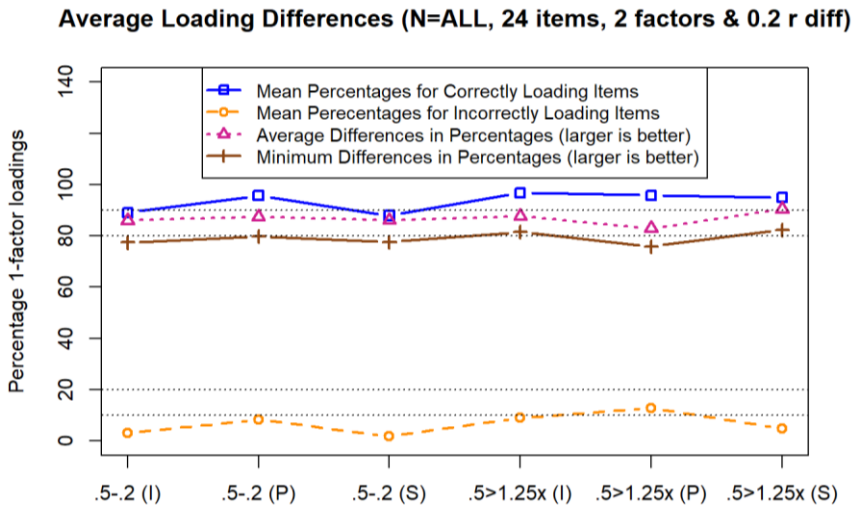
(a).



(b).

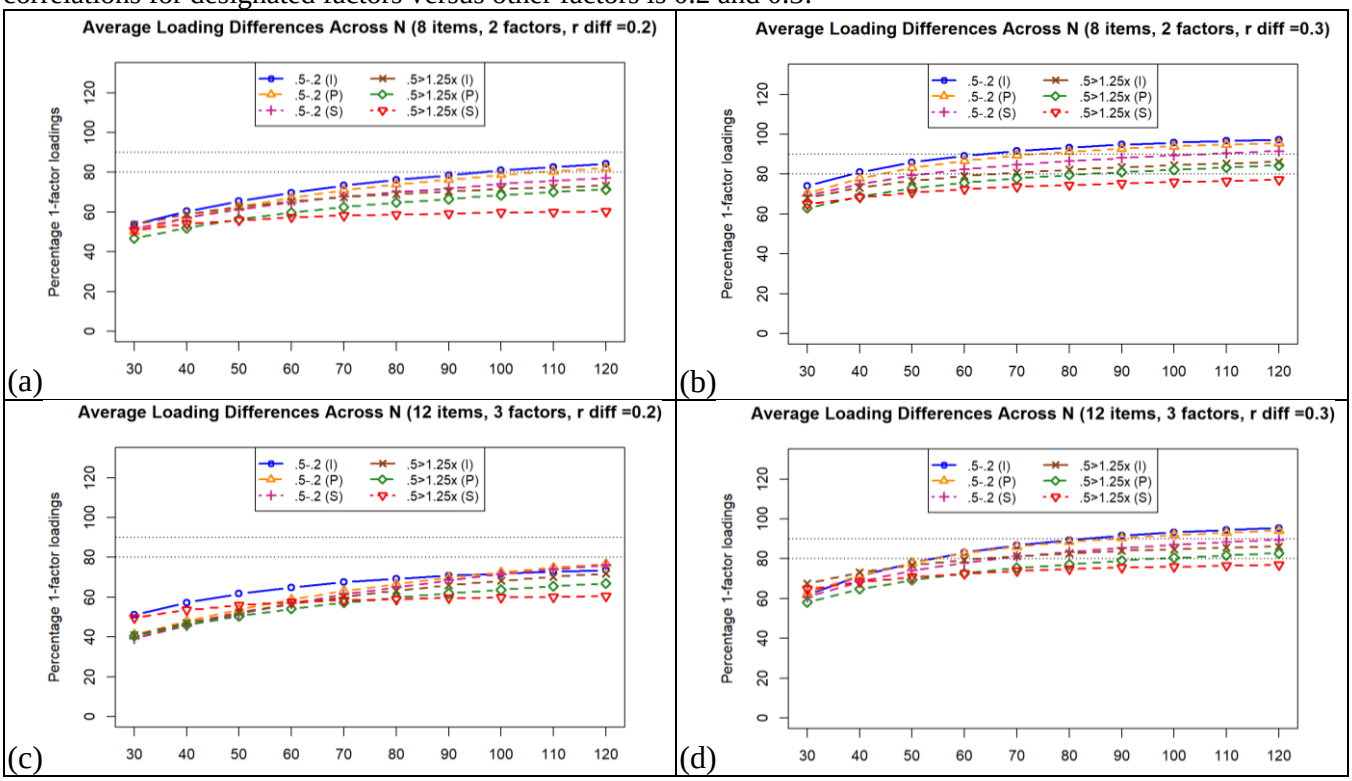


(c).



Note*: (I) represents SITCA; (P) represents PCA Loadings analysis using Pattern Matrix; (S) represents PCA Loadings analysis using Structure Matrix*

Figure 6. Comparing Average Differences between Loading Power percentages and Loading Error percentages for 4:1 Item-to-Factor Ratio conditions across sample sizes when the difference between correlations for designated factors versus other factors is 0.2 and 0.3.



Note*: (I) represents SITCA; (P) represents PCA Loadings analysis using Pattern Matrix; (S) represents PCA Loadings analysis using Structure Matrix*

any dimension to load, nor do we want cross-loaded items to load on only a single dimension. SITCA with the .5-.2 rule performs effectively across many conditions in this regard. The results in Figures 12-14 show that samples of about 50-60 may be sufficient in 6:1 item-to-factor (with r_diff = 0.2) conditions to identify not-loaded/cross-loaded items using SITCA with the .5-.2 rule (and with smaller samples for r_diff = 0.3).

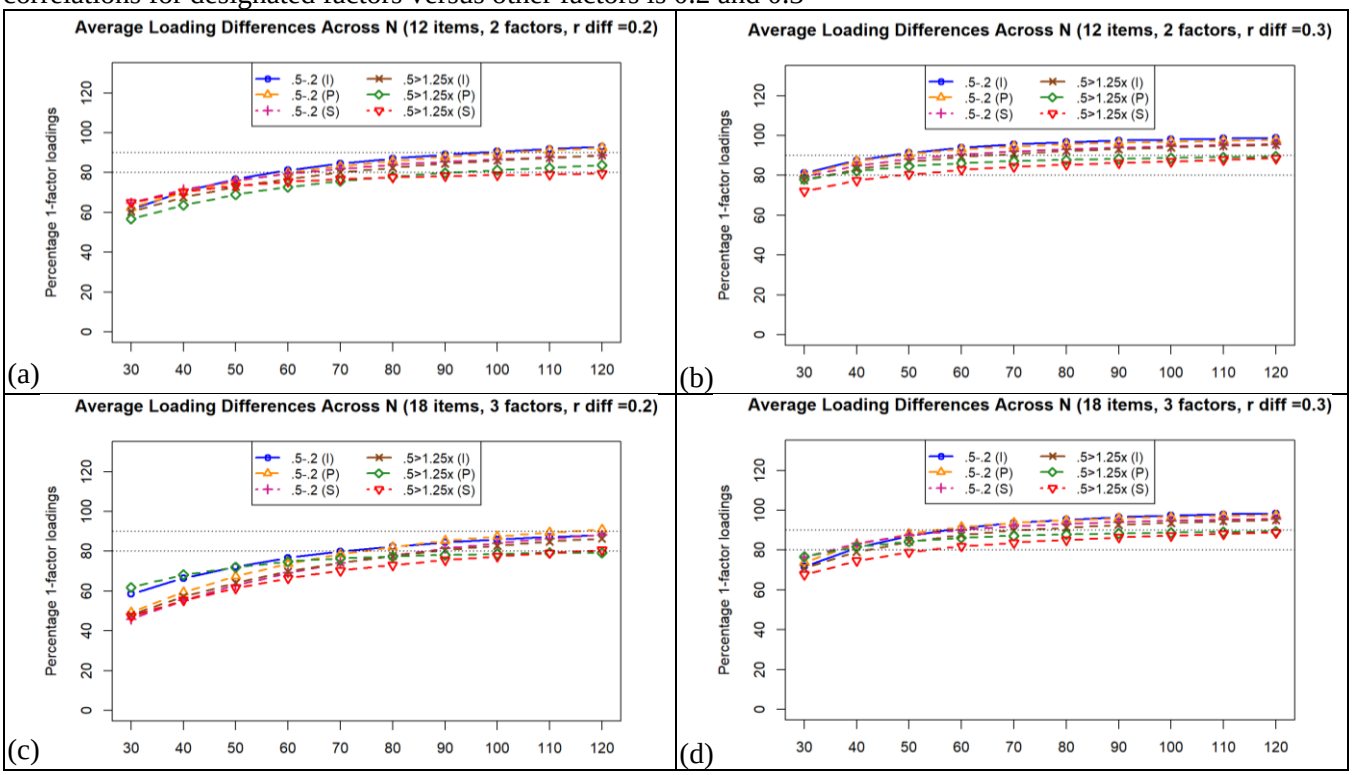
Conclusions

The purpose of this study was to investigate SITCA, compared to PCA, with smaller samples that might be more obtainable during small-scale studies. That is, while still creating and revising items, researchers want to diagnose whether items load on any dimensions of the scale, load on the correct dimensions, or load ideally on just one dimension. Diagnosing problematic items early in the scale development process while still revising items, with smaller samples, will allow developers to revise items and scales before expending time, money, and effort collecting larger samples for structural and construct validity evidence purposes.

We found that SITCA, especially combined with the .5-.2 simple structure rule, shows promise in its potential to identify not-loaded/cross-loaded items in scale development with smaller sample sizes (N = 50-60) than required for EFA, CFA, or PCA. No methods were effective under N = 70 with 4:1 item-to-factor ratios or when the population correlation differentials were 0.1, but SITCA was more effective than PCA methods for diagnosing loading errors. The purpose here is narrow: identifying problematic items because they cross-load on multiple subscales or do not load at all on any scale—not structural validity evidence.

Our goal is to improve items. For this purpose, we are less concerned with small sample results that correctly identify that items belong on particular subscale (loading power)—because we believe this falls more appropriately in the realm of structural validity evaluation. For item analysis, we believe the focus should be on minimizing loading errors. That is, lower loading error rates imply that not-loaded/cross-loaded items will be less likely to load incorrectly on a single subscale—and therefore, we can be more confident that items are genuinely concerns when they cross-load or do not load at all.

Figure 7. Comparing Average Differences between Loading Power percentages and Loading Error percentages for 6:1 Item-to-Factor Ratio conditions across sample sizes when the difference between correlations for designated factors versus other factors is 0.2 and 0.3



Note: *(I) represents SITCA; (P) represents PCA Loadings analysis using Pattern Matrix; (S) represents PCA Loadings analysis using Structure Matrix.*

We recommend that researchers follow strong processes when developing scales, such as phases suggested by various authors (Boateng et al., 2018; Clark & Watson, 1995; DeVellis, 2012; Worthington & Whittaker, 2006). Essentially these phases include:

1. defining the content domain for construct, including dimensions of the construct;
2. generating items, preferably within a table of item specifications;
3. evaluating content validity evidence;
4. evaluating items for both quality and bias, and choosing, repairing, removing and replacing items, as necessary;
5. evaluating scale reliability and construct validity evidence;
6. using the scale and continuing to evaluate and revise as needed.

Some have recommended a mixed methods approach to generating items in Phase 2 (Zhou, 2019), but others rely on theoretical definitions of constructs. If analyses during Phases 3-5, which typically require new data collection, suggest changes to the instrument, then the analyses in Phases 3-5 would ideally be repeated with new samples (but perhaps all analyses could be performed in one new larger validity study sample if changes are made in Phase 5). Along with item analyses (item difficulty and discrimination), the SITCA cross-loadings analysis presented in this paper would occur in Phase 4 for the purpose of evaluating items. Phase 4 can often be accomplished with small (but well-taken) samples, while Phase 5 is where the validity studies would begin and typically require larger samples. More complete EFA, CFA, or PCA would occur during Phase 5 to test reliability and structural validity to provide evidence of construct validity. SITCA cannot provide evidence for construct validity in Phase 5, but we would recommend that researchers repeat the item and cross-loadings analyses again in Phase 5—now with the benefit of larger samples (and maybe using EFA or PCA, or perhaps CFA, rather than SITCA).

Cross-loading can occur when subscale dimensions are not as differentiated as the researcher may believe (a more theoretical matter), which makes it difficult to justify that the dimensions represent separate concepts. These more theoretical cross-loadings are diagnosed through EFA, CFA, or PCA using large

Figure 8. Comparing Average Differences between Loading Power percentages and Loading Error percentages for 8:1 Item-to-Factor Ratio conditions across sample sizes when the difference between correlations for designated factors versus other factors is 0.2 and 0.3

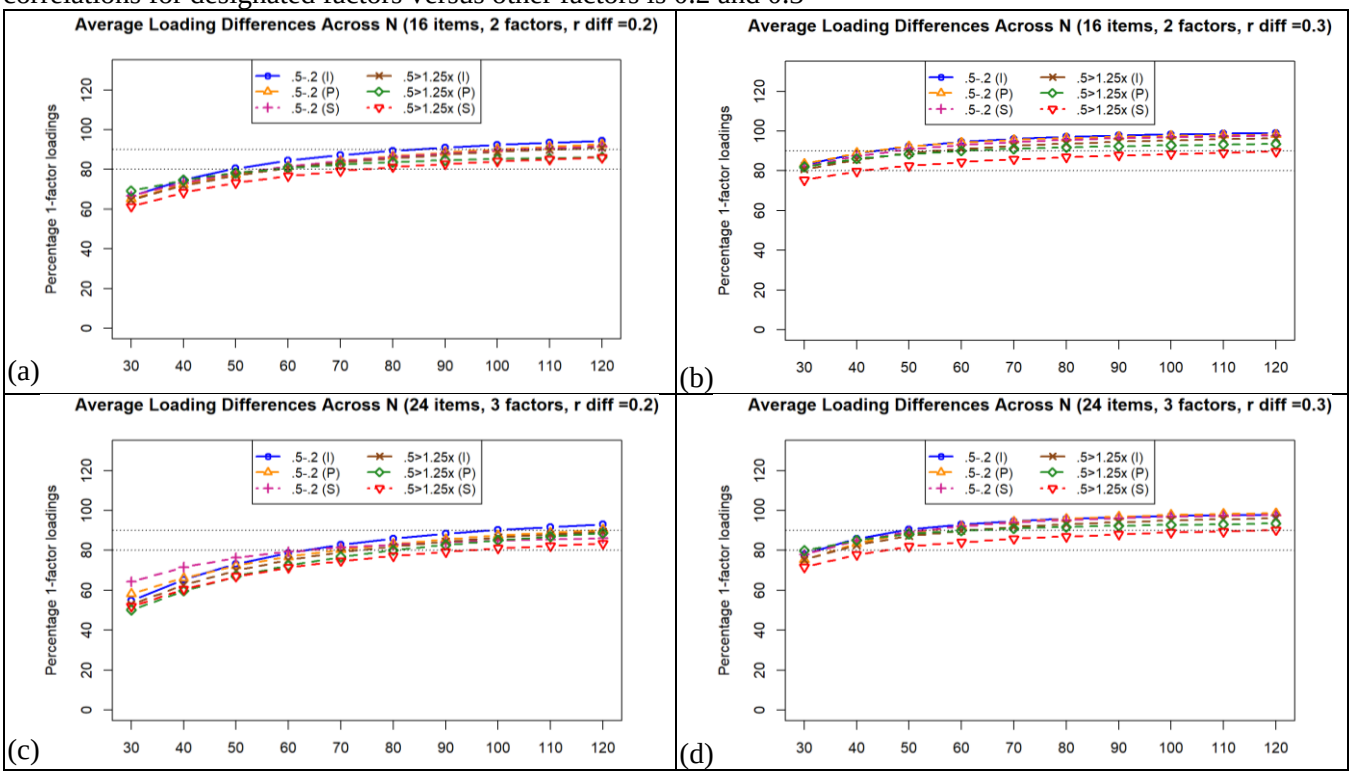
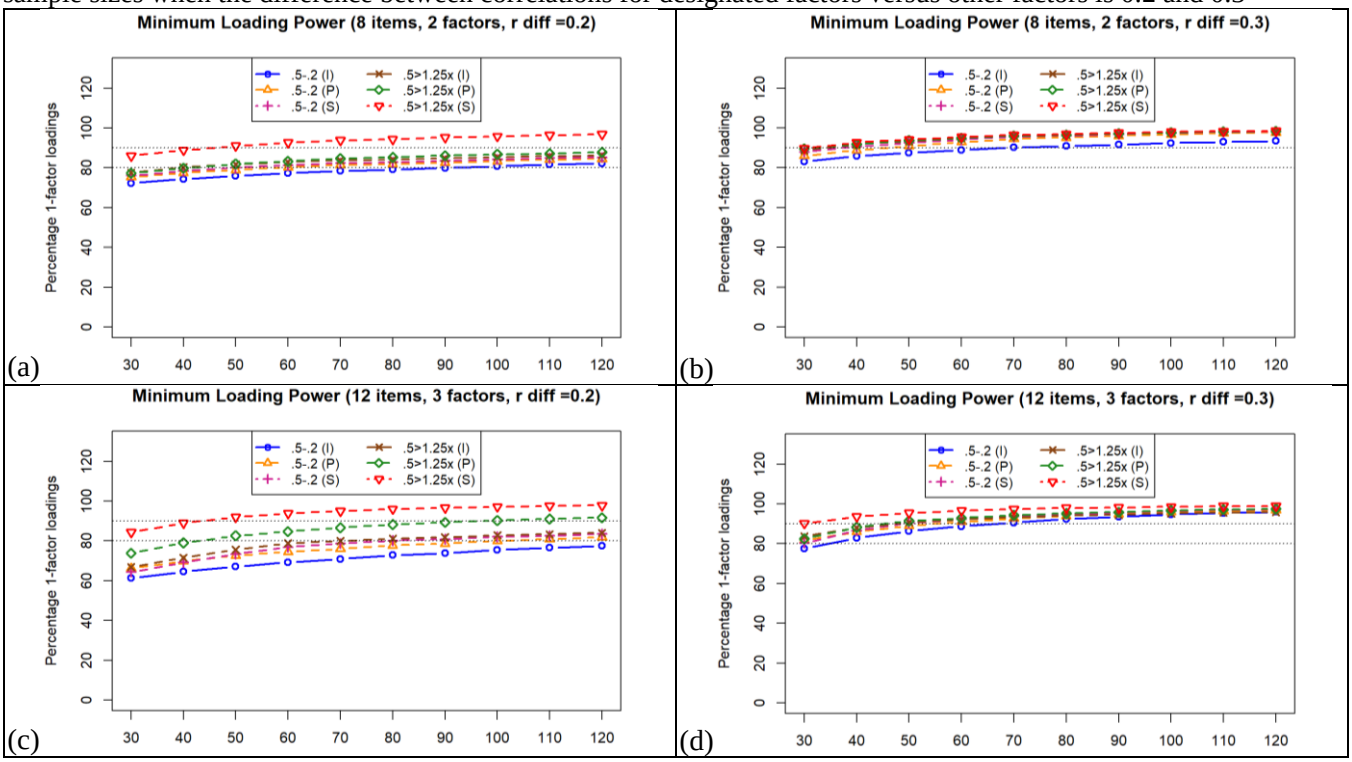
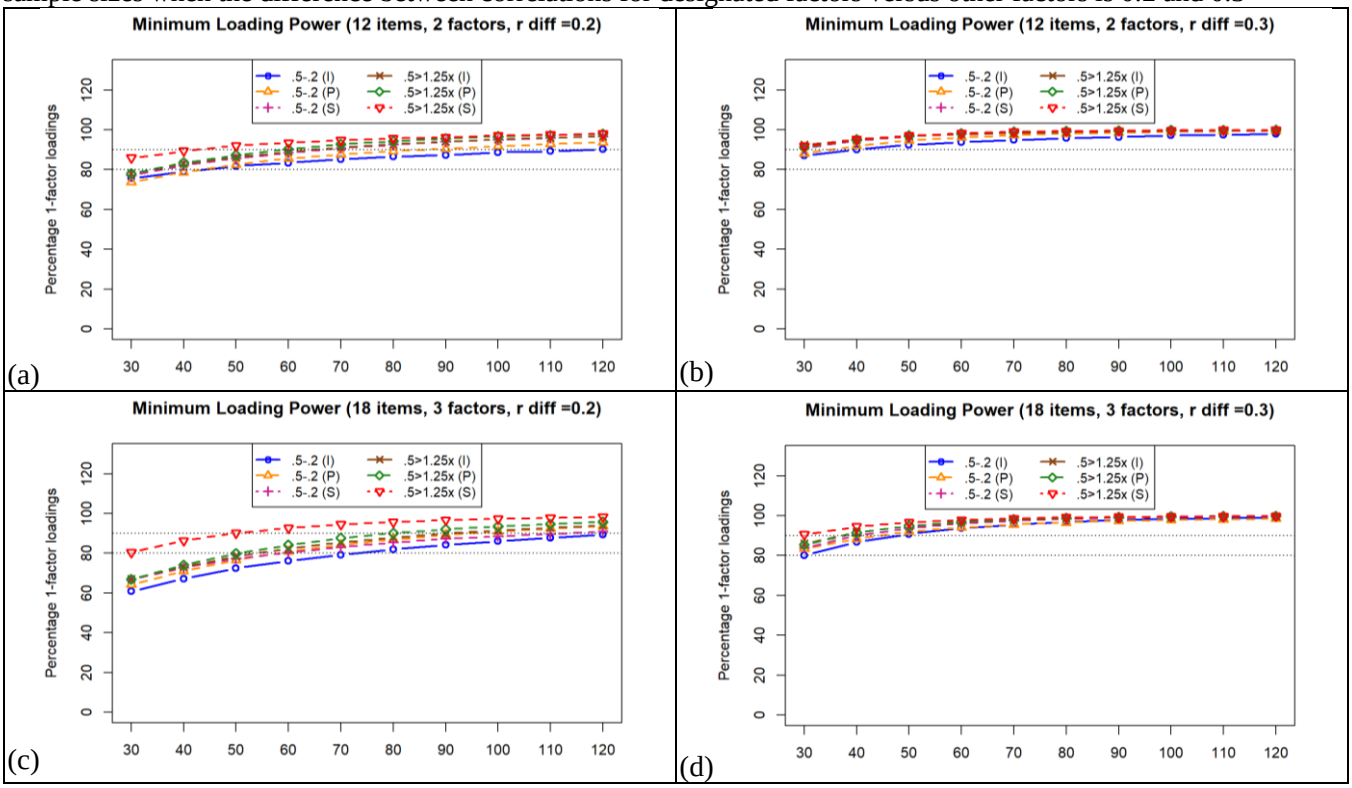


Figure 9. Comparing the Minimum Loading Power percentages for 4:1 Item-to-Factor ratio conditions across sample sizes when the difference between correlations for designated factors versus other factors is 0.2 and 0.3



Note*: (I) represents SITCA; (P) represents PCA Loadings analysis using Pattern Matrix; (S) represents PCA Loadings analysis using Structure Matrix.*

Figure 10. Comparing the Minimum Loading Power percentages for 6:1 Item-to-Factor ratio conditions across sample sizes when the difference between correlations for designated factors versus other factors is 0.2 and 0.3



Note*: (I) represents SITCA; (P) represents PCA Loadings analysis using Pattern Matrix; (S) represents PCA Loadings analysis using Structure Matrix.*

samples. Rotation can sometimes help remove these initial cross-loadings, but sometimes it is a deeper theoretical matter such as converging two dimensions into one rather than attempting to measure them separately. However, for cross-loadings that occur for more technical reasons, such as because of the wording of an item or other flaws in the item, we believe that SITCA with the .5-.2 simple structure rule can be used with smaller samples (N = 50-60) to diagnose such non-loading and cross-loading items. This recommendation is due primarily to its relatively lower loading errors. Indeed, SITCA with .5-.2 appears to work for this purpose more effectively than PCA approaches when there were at least six items per sub scale even with larger sample sizes (recall that none of the methods worked satisfactorily well with the smallest samples in the 4:1 ratio conditions when r_diff was 0.1 or less). It is also worth noting (though not shown) that with at least 8 items per factor, SITCA produced results somewhat better than other methods even with r_diff = 0.1, except that usually 20-30 more cases than r_diff = 0.2 were required to reach the 20% criterion. That is, with SITCA these problematic items rarely loaded incorrectly on a single dimension as compared to other methods. Scale developers can use this approach with some confidence to diagnose these types of problematic items by the fact that they did not load on just a single dimension (subscale).

One approach that scale developers can take at the piloting stage is to obtain slightly larger samples than the 24-36 that some scholars have recommended for item analysis pilot studies (Johanson & Brooks, 2010). By collecting data from 50-60 cases, researchers can perform both item analyses and SITCA cross-loading analyses with the same sample. Alternatively, scale developers might work toward stronger items before cross-loading analyses by using strategies that include small scale pilot studies for item analyses and perhaps by using mixed methods cognitive interviewing or think-aloud techniques. Stronger content validity and item analysis evidence before collecting data for cross-loading analyses may help create stronger items that will load more cleanly on the correct dimensions when the cross-loading investigation begins—thereby requiring smaller samples. Future research might examine corrected (or adjusted) item-total correlations used in this SITCA process, test additional simple structure rules, and whether Spearman or Polychoric correlations work more effectively than Pearson correlations (some authors have suggested Pearson correlations can be justified in the scenarios examined here, e.g., Robitzsch, 2020).

Figure 11. Comparing the Minimum Loading Power percentages for 8:1 Item-to-Factor ratio conditions across sample sizes when the difference between correlations for designated factors versus other factors is 0.2 and 0.3

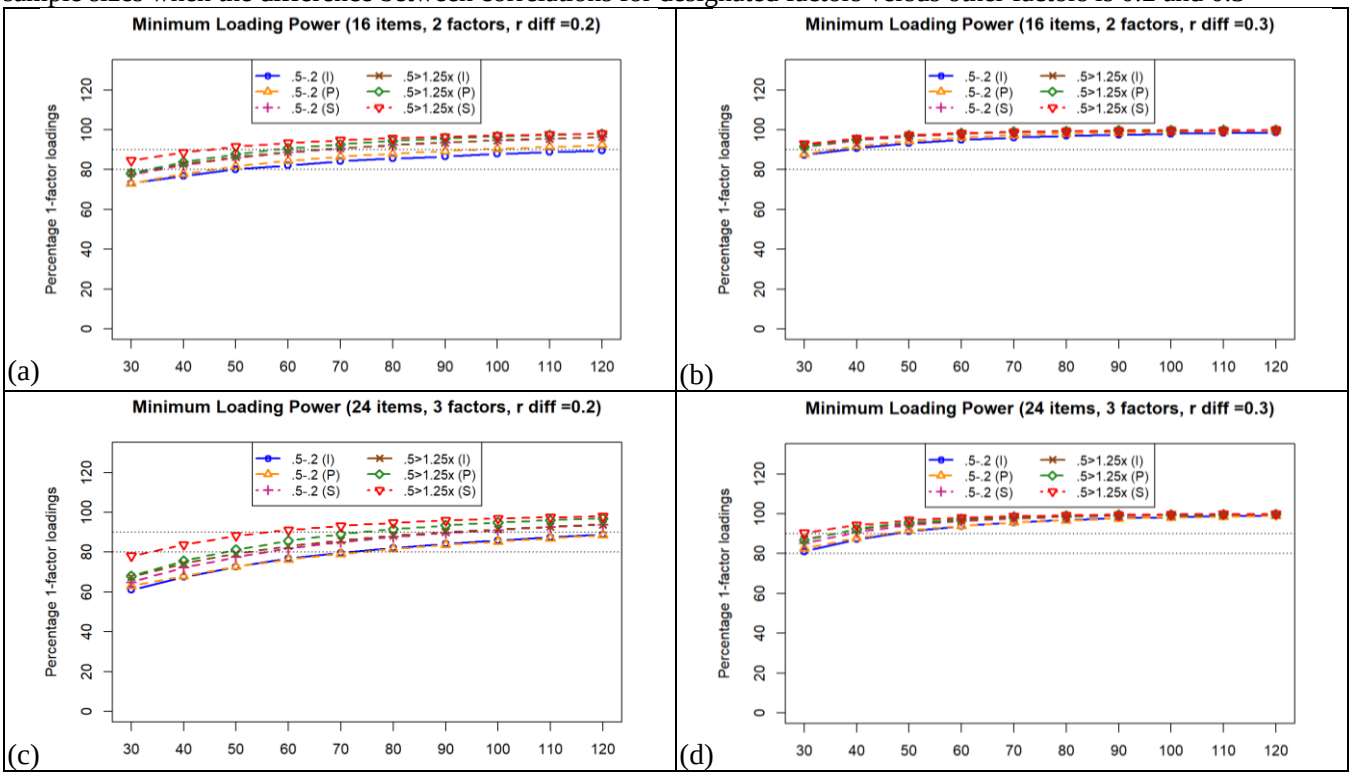
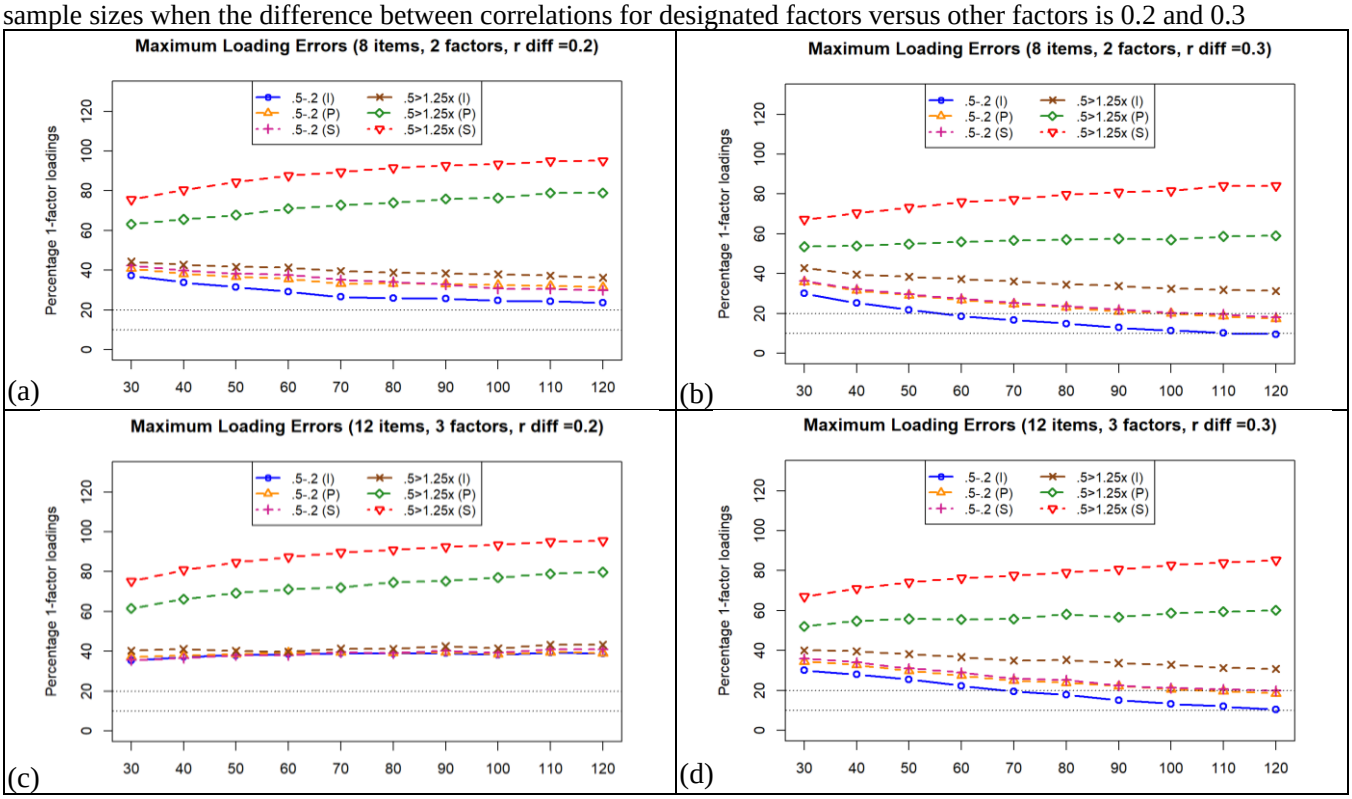


Figure 12. Comparing the Maximum Loading Error percentages for 4:1 Item-to-Factor ratio conditions across sample sizes when the difference between correlations for designated factors versus other factors is 0.2 and 0.3



Note: *(I) represents SITCA; (P) represents PCA Loadings analysis using Pattern Matrix; (S) represents PCA Loadings analysis using Structure Matrix*

Figure 13. Comparing the Maximum Loading Error percentages for 6:1 Item-to-Factor ratio conditions across sample sizes when the difference between correlations for designated factors versus other factors is 0.2 and 0.3

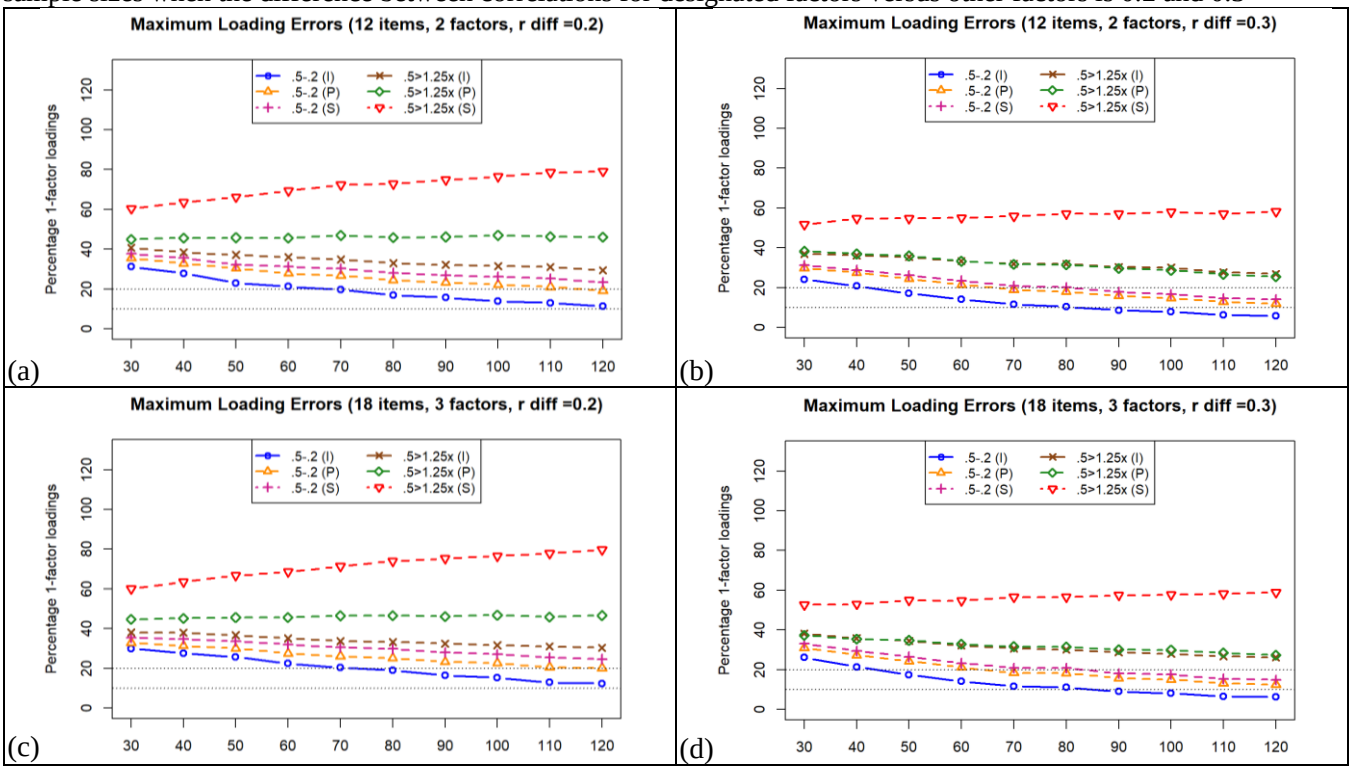
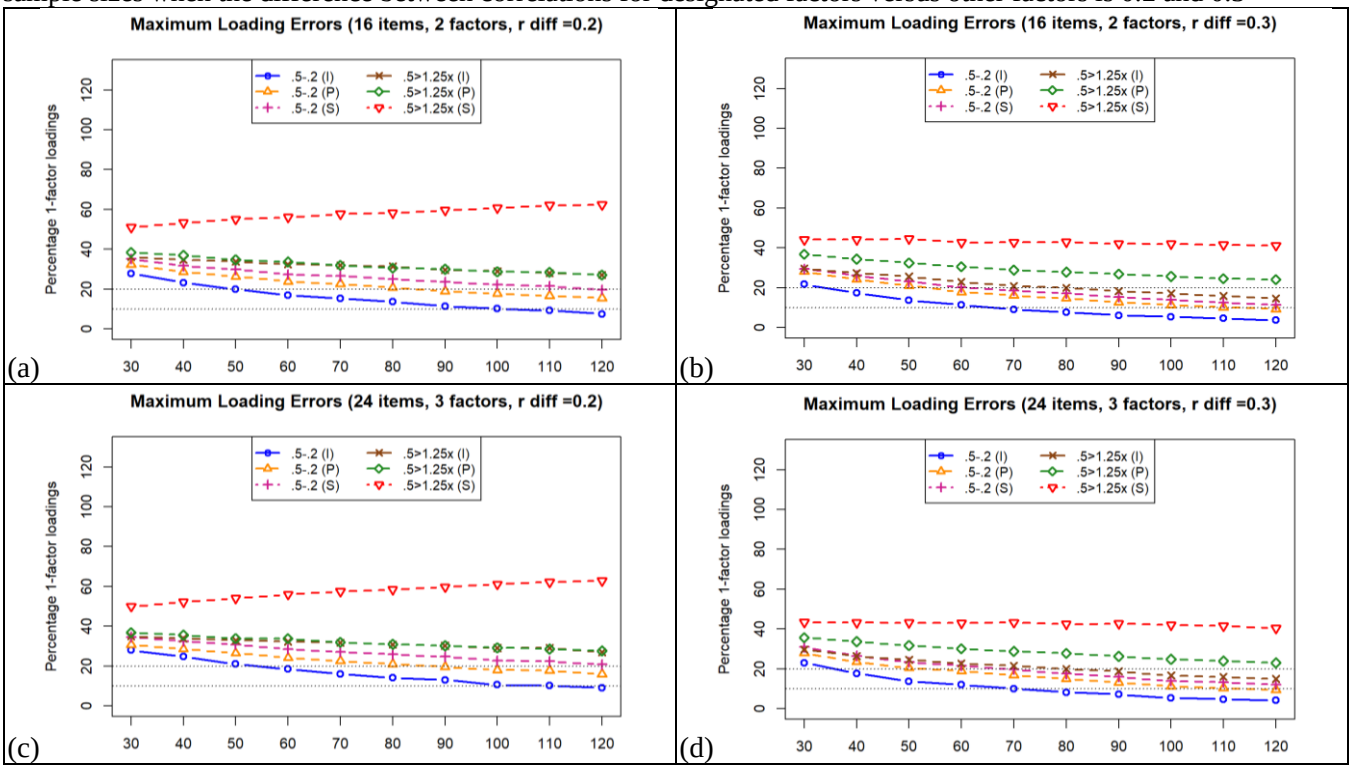


Figure 14. Comparing the Maximum Loading Error percentages for 8:1 Item-to-Factor ratio conditions across sample sizes when the difference between correlations for designated factors versus other factors is 0.2 and 0.3



Note*: (I) represents SITCA; (P) represents PCA Loadings analysis using Pattern Matrix; (S) represents PCA Loadings analysis using Structure Matrix*

References

Bai, J., & Ng, S. (2008). *Large dimensional factor analysis*. Now Publishers Inc.

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. Guilford Publications.

Bandalos, D. L., & Finney, S. J. (2010). Exploratory and confirmatory. *The reviewer's guide to quantitative methods in the social sciences*, 93.

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health, 6*, Article 149. DOI: 10.3389/fpubh.2018.00149

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*(3), 309-319.

DeVellis, R. F. (2012). *Scale development: Theory and applications* (3rd ed.). Sage.

de Winter, J. C. F., Dodou, D., & Wieringa, P. A. (2009). Exploratory factor analysis with small sample sizes. *Multivariate Behavioral Research, 44*, 147–181. DOI: 10.1080/00273170902794206

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (1998). *Multivariate data analysis* (Vol. 5, No. 3, pp. 207-219). Upper Saddle River, NJ: Prentice Hall.

Henrysson, S. (1962). The relation between factor loadings and biserial correlations in item analysis. *Psychometrika, 27*(4), 419-424.

Guadagnoli, E., & Velicer, W. F. (1988). Relation to sample size to the stability of component patterns. *Psychological Bulletin, 103*(2), 265-275. DOI: 10.1037/0033-2909.103.2.265

Johanson, G. A., & Brooks, G. P. (2010). Initial scale development: Sample size for pilot studies. *Educational and Psychological Measurement, 70*(3), 394-400. DOI: 10.1177/0013164409355692

Jordan, P., & Spiess, M. (2019). Rethinking the interpretation of item discrimination and factor loadings. *Educational and Psychological Measurement, 79*(6), 1103-1132. DOI:10.1177/0013164419843164

Kline, P. (1994). *An easy guide to factor analysis*. Routledge.

Kline, R. (2016). Principles and Practice of Structural Equation Modeling (4th ed.). Guilford.Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum Sample Size Recommendations for Conducting Factor Analyses. *International Journal of Testing, 5*(2), 159–168.

Pearson, R. H., & Mundfrom, D. J. (2010). Recommended Sample Size for Conducting Exploratory Factor Analysis on Dichotomous Data. *Journal of Modern Applied Statistical Methods, 9*(2), Article 5.

Pituch, K. A., & Stevens, J. P. (2015). *Applied multivariate statistics for the social sciences: Analyses with SAS and IBM's SPSS*. Routledge.

Richardson, M. W. (1936). Notes on the rationale of item analysis. *Psychometrika, 1*(1), 69-76.

Robitzsch, A. (2020). Why Ordinal Variables Can (Almost) Always Be Treated as Continuous Variables: Clarifying Assumptions of Robust Continuous and Ordinal Factor Analysis Estimation Methods, *Frontiers in Education, 5*, Article 589965. DOI: 10.3389/feduc.2020.589965

Rummel, R.J. (1970). *Applied factor analysis*. Evanston, IL: Northwestern University Press.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Allyn & Bacon

Thurstone, L. (1947). The simple structure concept. *Multiple Factor Analysis: A Development and Expansion of The Vectors of Mind*, 319-346.

Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research, 25*(1), 1-28.

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist, 34*, 806-838. DOI: 10.1177/0011000006288127

Zhou, Y. (2019). A mixed methods model of scale development and validation analysis. *Measurement: Interdisciplinary Research and Perspectives, 17*(1), 38-47. DOI: 10.1080/15366367.2018.1479088

Send correspondence to: Gordon Brooks
Ohio University
Email: brooksg@ohio.edu