Finally! A valid test of configural invariance using permutation in multigroup CFA

Jorgensen, T.D.; Kite, B.A.; Chen, P.-Y.; Short, S.D.

DOI
10.1007/978-3-319-56294-0_9

Publication date
2017

Document Version
Final published version

Published in
Quantitative psychology

Finally! A Valid Test of Configural Invariance Using Permutation in Multigroup CFA

Terrence D. Jorgensen, Benjamin A. Kite, Po-Yi Chen, and Stephen D. Short

Abstract In multigroup factor analysis, configural measurement invariance is accepted as tenable when researchers either (a) fail to reject the null hypothesis of exact fit using a χ^2 test or (b) conclude that a model fits approximately well enough, according to one or more alternative fit indices (AFIs). These criteria fail for two reasons. First, the test of perfect fit confounds model fit with group equivalence, so rejecting the null hypothesis of perfect fit does not imply that the null hypothesis of configural invariance should be rejected. Second, treating common rules of thumb as critical values for judging approximate fit yields inconsistent results across conditions because fixed cutoffs ignore sampling variability of AFIs. As a solution, we propose replacing χ^2 and fixed AFI cutoffs with permutation tests. Iterative permutation of group assignment yields an empirical distribution of any fit measure under the null hypothesis of invariance. Simulations show the permutation test of configural invariance controls Type I error rates better than χ^2 or AFIs when a model has parsimony error (i.e., negligible misspecification) but the factor structure is equivalent across groups (i.e., the null hypothesis is true).

Keywords Measurement equivalence • Configural invariance • Permutation • Multiple group confirmatory factor analysis

1 Introduction

The assumption of measurement equivalence/invariance (ME/I) is required to draw inferences about how latent constructs might differ across different contexts, such as different occasions or populations. Configural ME/I, in particular, must

T.D. Jorgensen (✉)
University of Amsterdam, Amsterdam, The Netherlands
e-mail: T.D.Jorgensen@uva.nl

B.A. Kite • P.-Y. Chen
University of Kansas, Lawrence, KS 66045, USA

S.D. Short
College of Charleston, Charleston, SC 29424, USA

L.A. van der Ark et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 196, DOI 10.1007/978-3-319-56294-0_9

be implicitly assumed before measurement parameters can be compared across contexts, whose equality is required before comparing common-factor parameters across contexts. Multigroup confirmatory factor analysis (CFA) is one of the most common frameworks used to test ME/I across groups (Vandenberg and Lance 2000), and multigroup models provided the only avenue for testing whether factor structure is configured equivalently across groups.

We first describe the current recommended best practices for testing configural invariance, as well as their limitations. We then propose a permutation randomization test of configural invariance across groups. We present Monte Carlo simulation studies to compare the power and Type I error rates of the permutation method to other methods.

1.1 Assessing Configural Invariance

To test for configural invariance (i.e., the same form), researchers fit a model with identical factor structure across groups, but allow all freely estimated measurement-model parameters (factor loadings, intercepts, and residual variances) to differ between groups (except scale-identification constraints). The likelihood ratio test (LRT or χ^2 statistic of exact fit) is used to judge whether the configural invariance model is an acceptable baseline model before constraining measurement parameters (Byrne et al. 1989). If the test is not significant at the specified α level, the analyst fails to reject the null hypothesis (H_0) that the configural model fits well and proceeds to test equality of item parameters (e.g., factor loadings, intercepts or thresholds, residual variances) across groups.

The LRT confounds two sources of model misfit (Cudeck and Henly 1991; MacCallum 2003): estimation discrepancy (due to sampling error) and approximation discrepancy (due to a lack of correspondence between the population and analysis models). Because configural ME/I is assessed by testing the absolute fit of the configural model, an LRT for a multigroup model further confounds two sources of approximation discrepancy. The overall lack of correspondence between the population and analysis models could theoretically be partitioned into (a) differences among the groups' true population models and (b) discrepancies between each group's population and analysis models. It is possible (perhaps even probable) that an analysis model corresponds only approximately to the groups' population models (Byrne et al. 1989), yet the analysis model may be equally (in)appropriate for each group. Although overall model fit is certainly important to assess in conjunction with tests of ME/I, the H_0 of configural invariance is only concerned with group equivalence, so the LRT does not truly provide a test of configural invariance.

Large sample sizes make the LRT sensitive even to minute differences in model form, which would have little or no practical consequence on parameter estimates. Many researchers would prefer to use an alternative fit index (AFI) to assess the approximate fit of the configural model. Putnick and Bornstein (2016) found that

only 17% of ME/I tests are decided by the LRT alone, whereas 46% also involve at least one AFI, and 34% are decided using AFIs alone. The comparative fit index (CFI) (Bentler 1990) was reported for 73.2% of ME/I tests, making it the most popular AFI in this context. This chapter will focus mainly on CFI, but the root mean square error of approximation (RMSEA) (Steiger and Lind 1980) is also very popular. AFIs are functions of overall discrepancies between observed and model-implied sample moments, so using them to assess configural invariance would confound group equivalence with overall misfit, just like the LRT. However, we discuss their additional limitations below.

Most AFIs do not have known sampling distributions,[1] so evaluating the fit of a configural model involves some subjective decisions (e.g., which fit indices to use, what values indicate acceptable fit). Sometimes there are conflicting recommendations based on different criteria. For example, Bentler and Bonett (1980) suggested CFI > 0.90 indicates good fit, yet Hu and Bentler (1999) recommended CFI > 0.95 as a stricter criterion. Browne and Cudeck (1992) suggested RMSEA < 0.05 indicates close fit, RMSEA < 0.08 indicates reasonable fit, and RMSEA > 0.10 indices poor fit (RMSEA between 0.08 and 0.10 indicates mediocre fit) (MacCallum et al. 1996), yet Hu and Bentler recommended RMSEA < 0.06 as a stricter criterion. According to an October 2016 Google Scholar search, Hu and Bentler's criteria seem to be more widely applied (35,474 citations) than Bentler and Bonett's (13,815 citations) or Browne and Cudeck's (2843 citations).

The problem with using fixed cutoffs, even as mere rules of thumb, is that they ignore conditions specific to the study, such as sample size (and by implication, sampling error), number of groups, sample size ratios, number (and pattern) of indicators and factors, etc. Fixed cutoffs can also lead to the apparent paradox that larger samples yield lower power, which occurs when the AFI cutoff is more extreme than the population-level AFI[2] (Marsh et al. 2004).

1.2 A Permutation Randomization Test of Configural Invariance

When a theoretical distribution of a statistic is unavailable for null-hypothesis significance testing, it is possible for researchers to use a resampling method to create an empirical sampling distribution from their observed data. Rodgers (1999) provided a useful taxonomy of resampling methods. One flexible method that can be used to create an empirical approximation of a sampling distribution is the permutation randomization test. If a method of resampling the data can be conceived such that a H_0 is known to be true (in the permutation distribution),

[1] A notable exception is RMSEA. See an excellent discussion by Kenny et al. (2015).

[2] Population-level AFIs can be obtained by fitting the analysis model to the population moments or can be estimated from the average AFI across Monte Carlo samples.

then reference distributions can be empirically approximated for statistics whose sampling distributions are unknown or intractable.

The logic of the permutation test is related to the use of random assignment in experimental designs. Random assignment of subjects to two (or more) groups will average out any between-group differences, so that on average, group mean differences would be zero, resulting in two (or more) comparable groups before administering different treatments. Due to sampling fluctuation, observed differences will not be exactly zero after any single random assignment, but differences will be zero on average across replications of random assignment. Capitalizing on this effect of randomization, when a set of observed outcome scores (Y) is randomly (re)assigned to the two different observed groups (natural[3] or experimental), any existing between-group differences would be zero, on average.

A simple example of a permutation test is to compare two group means. The grouping variable (G) can be resampled without replacement and paired with values on the dependent variable (Y). The resulting randomization is a single permutation (reordering) of the data. Because H_0: $\mu_1 - \mu_2 = 0$ is true (i.e., the groups do not systematically differ in a permuted data set), the calculated t value is one observation from a theoretically infinite population of t statistics that could be calculated under the H_0 of no group mean difference. Repeating this process 100 times results in a distribution of 100 t statistics under H_0, one t value from each permutation of the data. As the number of permutations increases, the shape of the empirical distribution of the t values will become a closer approximation of the true, but unknown, sampling distribution. Using the empirical approximation of the sampling distribution under H_0, a researcher can calculate a good approximate p value by determining the proportion of the permutation distribution that is more extreme than the t value calculated from the original, unpermuted data.

We propose a permutation method for testing configural invariance. Randomly permuting group assignment yields resampled data for which the H_0 of group equivalence in model fit is true, even if the model does not fit perfectly. The steps to test configural ME/I are:

1. Fit the hypothesized multiple-group model to the original data, and save the fit measure(s) of interest.
2. Sample N values without replacement from the observed grouping-variable vector G. The new vector $G_{\text{perm(i)}}$ contains the same values as G, but in a new randomly determined order (i.e., $G_{\text{perm(i)}}$ is a permutation of G).
3. Assign the nth row of original data to the nth permuted value from $G_{\text{perm(i)}}$. On average, group differences are removed from this ith permuted data set.
4. Fit the same multiple-group model from step 1 to the permuted data, and save the same fit measure(s).

[3]The exchangeability assumption might be violated for natural groups (Hayes 1996), which we bring up in the Discussion.

5. Repeat steps 2–4 *I* times, resulting in a vector of length *I* for each fit measure.
6. Make an inference about the observed fit measure by comparing it to the vector of permuted fit measures.

Step 6 can test H_0 in either of two ways, yielding the same decision:

- Calculate the proportion of the vector of permuted fit measures that is more extreme (i.e., worse fit) than the observed fit measure. This is a one-tailed *p* value that approximates the probability of obtaining a fit measure at least as poor as the observed one, if the H_0 of ME/I for all groups holds true. Reject H_0 if $p < \alpha$.
- Sort the vector of permuted fit measures in ascending order for badness of fit measures like χ^2 or RMSEA or sort in descending order for goodness of fit indices like CFI. Use the $[100 \times (1 - \alpha)]$th percentile as a critical value, and reject H_0 if the observed fit measure is more extreme than the critical value.

Because permutation removes group differences (on average) without altering the structure among the variables in any other way, this method provides a simple framework to test configural ME/I separately from overall model fit. Furthermore, permutation provides empirical sampling distributions of AFIs, which generally have unknown sampling distributions. Researchers using permutation methods would not need to rely on fixed cutoff criteria based on intuition or studies whose simulated conditions might not closely resemble their own data and model(s), such as CFI > 0.90 (Bentler and Bonett 1980) or CFI > 0.95 (Hu and Bentler 1999). As we demonstrate using simulation studies, none of these fixed rules-of-thumb consistently control Type I error rates. In contrast, permutation distributions implicitly take into account the unique qualities of the data and model under consideration.

2 Monte Carlo Simulations

To evaluate the permutation methods proposed in the previous section, we present results from two simulation studies. The first evaluated Type I error rates when H_0 is true and second evaluated power when H_0 is false. In each study, we compared H_0 rejection rates between permutation methods and currently recommended practices under a variety of sample-size and model-size conditions.

We chose conditions based on Meade et al. (2008) ME/I study, which included approximation error in the form of near-zero cross-loadings in the population models that were fixed to zero in the analysis models (i.e., simple structure was only approximately true; see Table 1). When fitting simple structure CFA models to the model-implied population moments, AFIs indicated good model fit using standard conventions (e.g., CFI > 0.98, RMSEA < 0.03).

Based on Meade et al. (2008), we varied sample size (*N*) in each of two groups across five levels, 100, 200, 400, 800, and 1600 per group. We varied model complexity via number of factors (2 or 4) and number of items per factor (4 or 8), using the same population values for factor loadings as Meade et al. (2008) (see

Table 1 Population factor loadings (Λ matrix)

Item	Factor 1	Factor 2	Factor 3	Factor 4
1	0.68 (0.54)	−0.03	−0.02	−0.11
2	0.76 (0.62)	0.02	−0.03	−0.03
3	0.74 (0.60)	−0.04	−0.03	0.00
4	0.75 (0.61)	0.00	−0.01	0.08
5	0.04	0.76 (0.61)	0.07	0.00
6	−0.06	0.56 (0.41)	−0.03	0.04
7	−0.08	0.75 (0.60)	0.07	0.06
8	−0.02	0.72 (0.57)	0.05	−0.03
9	0.07	−0.01	0.80 (0.65)	0.00
10	−0.01	−0.03	0.58 (0.43)	−0.02
11	−0.04	0.06	0.80 (0.65)	0.03
12	0.04	0.00	0.39 (0.24)	0.05
13	−0.02	−0.02	−0.01	0.65 (0.51)
14	0.00	−0.13	−0.03	0.67 (0.53)
15	0.00	0.03	−0.01	0.59 (0.45)
16	0.00	0.03	0.02	0.67 (0.53)

Note. For conditions with eight indicators per factor, λ_s in parentheses were used as population parameters, and λ_s for items 17–32 were identical to λ_s for items 1–16. Cells with only one value (near zero) are minor discrepancies from simple structure (approximation error)

Table 1). In the population models, we fixed all intercepts to zero, discussed next), factor means to zero, factor variances to one, factor correlations to 0.3, and residual variances to values that would set total item variances to one (i.e., standard normal variables). We simulated 2000 replications in each condition and used $I = 200$ permutations to calculate p values associated with fit measures.

Whereas Meade et al. (2008) simulated configural lack of invariance (LOI) by adding additional factors to group two's population model (resulting in dozens of different population models), we simply changed one, two, three, or four of the zero (or nonsalient) parameters in group two's population model. The first level of configural LOI was to change factor loading λ_{51} from 0.04 to 0.7. The second level was to make the same change to λ_{51} and to add a residual covariance ($\theta_{72} = 0.2$). The third level made the same additions and changed λ_{12} from −0.03 to 0.7, and the fourth level also added another residual covariance ($\theta_{84} = 0.2$). These arbitrary levels of configural LOI served to compare the power of different methods to detect the same lack of correspondence between the groups' population models.

We used R (R Core Team 2016) to generate multivariate normal data and the R package lavaan (version 0.5–20; Rosseel 2012) to fit models to simulated data. Using lavaan's default settings, the scales of the latent factors were identified by fixing the first factor loading to one, and the latent means were fixed to zero.

3 Results

We first used a 5 (N) × 2 (2 or 4 factors) × 2 (4 or 8 items per factor) design, holding LOI constant at zero. Because the population models included minor approximation discrepancy in the form of near-zero cross-loadings, we expected Type I error rates to exceed 5% and for these rates to increase with N. Because fixed cutoffs do not take sampling variability or model complexity into account, we expected results to vary across Ns and model sizes. We expected permutation to yield nominal Type I error rates in all conditions for all fit measures, which would indicate a valid test of configural invariance.

As expected, using the traditional LRT resulted in extremely high Type I error rates. Figure 1 confirms that even in the condition with the smallest N and model, Type I errors were almost 20%, approaching 100% as N increased. For larger Ns, rejection rates matched the expected power using the Satorra and Saris (1985) method, but rejection rates were inflated at smaller N, especially in larger models, due to the small-sample bias of the LRT (Nevitt and Hancock 2004). In contrast, permutation provided nominal Type I error rates across conditions.

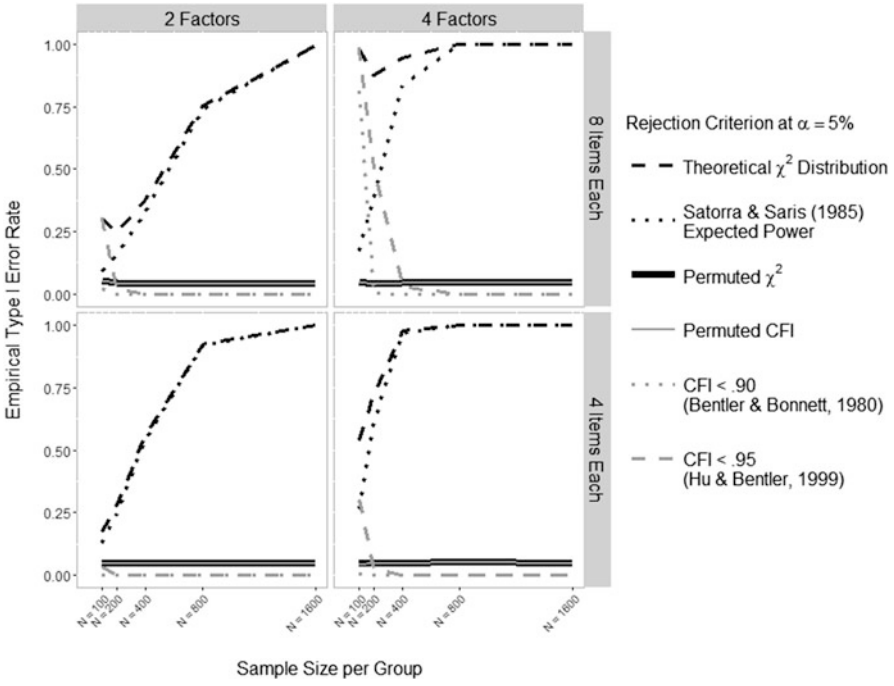

Fig. 1 Observed Type I error rates for LRT, CFI rules of thumb, and permutation tests of configural invariance, as well as expected power of LRT using the Satorra and Saris (1985) method

Using AFIs to assess approximate fit of the configural model only appeared to yield inflated Type I errors under small-*N* conditions, but that depended heavily on the size of the model and on which rule of thumb was used. Figure 1 shows that larger models yielded more errors at smaller *N*. Similar results were found for RMSEA guidelines. In contrast, permuting CFI (or any AFI) maintained nominal Type I error rates across all conditions.

We next used a 5 (*N*) × 4 (LOI) design, holding model complexity constant (4 items for each of 2 factors, the condition in which fixed cutoffs for CFI showed ≤5% Type I errors). We expected permutation to have lower power than the LRT, which already had high rejection rates when H_0 was true. Given that Type I error rates for AFI cutoffs were typically close to zero for this particular population model, we had no specific hypotheses about how their power would compare to power using permutation, but we did expect lower power with increasing *N* in conditions where population AFIs met guidelines for acceptable fit.

Figure 2 confirms our expectation that the LRT had the highest power to detect LOI, particularly at the lowest level of LOI and the smallest *N*. But as Fig. 1 shows, the greater power came at the expense of high Type I errors because the LRT tests overall model fit rather than configural invariance alone. Hu and Bentler's (1999) more stringent criterion (CFI > 0.95) yielded power almost as high as the LRT,

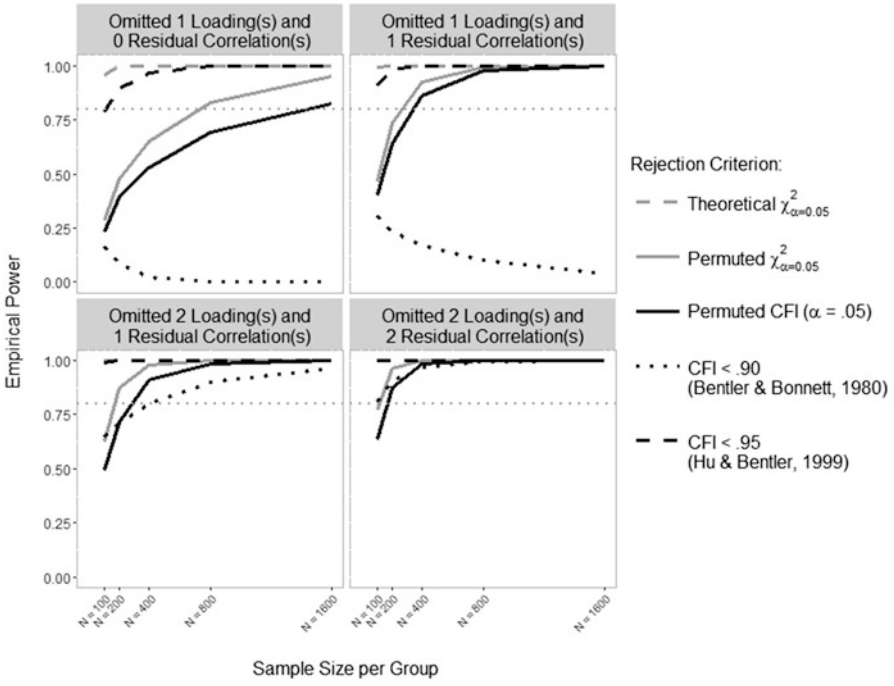

Fig. 2 Power for LRT (gray lines) and CFI (*black lines*) using theoretical (or fixed) vs. permutation-based critical values. The *dotted gray line* indicates 80% power

whereas Bentler and Bonett's (1980) less stringent criterion (CFI > 0.90) yielded lower power that decreased as N increased in conditions where only one or two salient population parameters differed between groups. We found the same pattern of results for RMSEA.

Permutation yielded inadequate power when only a single parameter differed between populations, unless $N \geq 800$ per group. Adequate power to detect greater LOI was achieved at smaller N. The permuted LRT tended to have greater power than permuted CFI, but the discrepancy was small when N and LOI were large. Permuted RMSEA had power similar to the permuted LRT.

4 Discussion

We proposed a permutation randomization framework for using multigroup CFA to test ME/I. We proposed this framework to address some limitations of current best practices. First, the LRT of exact (or equal) fit does not test the correct H_0 of group equivalence for the configural model. Assessing overall model fit confounds any group differences with overall model misspecification. Irrespective of how well a model only approximates a population process, the model may be equally well specified for both groups, in which case the H_0 of group equivalence should not be rejected. Our simulation studies showed that current best practices can lead to highly inflated Type I error rates, even for models with very good approximate fit. Permutation, on the other hand, yields well-controlled Type I error rates even when the model does not fit perfectly, providing the only valid test of configural invariance across groups that we are currently aware of.

Second, most researchers prefer AFIs over the LRT (Putnick and Bornstein 2016) because of the latter's sensitivity to differences that are negligible in practice, which could be thought of as inflated Type I error rates when assessing approximate fit in large samples. However, lack of known distributions for ΔAFIs leads to reliance on rule-of-thumb cutoffs that, as we have shown, lead to inflated Type I error rates in smaller (albeit still large) samples, especially in larger models. Our simulations showed that regardless of which fit measure is preferred, permutation provides well controlled Type I error rates, with power to detect true differences that is comparable to the LRT.

We recommend that applied researchers interested in testing configural invariance use the permutation method, which is implemented in a function called "`permuteMeasEq()`" in the R package `semTools` (semTools Contributors 2016). If the overall fit of the configural model is satisfactory, the permutation method provides a valid test of the H_0 of group equivalence in model form and is currently the only method to do so. Permutation may be particularly valuable in conditions with inflated error rates, such as missing or categorical data, but its utility may be limited by the exchangeability assumption. We encourage further investigation of permutation methods for testing group equivalence, particularly

for developing guidelines for modifying individual group models (when configural invariance does not hold) versus making modifications to poorly fitting models simultaneously across groups (when configural invariance does hold).

References

P.M. Bentler, Comparative fit indexes in structural models. Psychol. Bull. **107**(2), 238–246 (1990). doi:10.1037/0033-2909.107.2.238

P.M. Bentler, D.G. Bonett, Significance tests and goodness of fit in the analysis of covariance structures. Psychol. Bull. **88**(3), 588–606 (1980). doi:10.1037/0033-2909.88.3.588

M.W. Browne, R. Cudeck, Alternative ways of assessing model fit. Sociol. Methods Res. **21**, 230–258 (1992). doi:10.1177/0049124192021002005

B.M. Byrne, R.J. Shavelson, B. Muthén, Testing for the equivalence of factor co-variance and mean structures: The issue of partial measurement invariance. Psychol. Bull. **105**(3), 456–466 (1989). doi:10.1037/0033-2909.105.3.456

R. Cudeck, S.J. Henly, Model selection in covariance structures analysis and the "problem" of sample size: a clarification. Psychol. Bull. **109**(3), 512–519 (1991). doi:10.1037//0033-2909.109.3.512

A.F. Hayes, Permutation test is not distribution-free: Testing H0: $\rho = 0$. Psychol. Methods **1**(2), 184–198 (1996). doi:10.1037/1082-989X.1.2.184

L.-T. Hu, P.M. Bentler, Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct. Equ. Model. **6**(1), 1–55 (1999). doi:10.1080/10705519909540118

D.A. Kenny, B. Kaniskan, D.B. McCoach, The performance of RMSEA in models with small degrees of freedom. Sociol. Methods Res. **44**(3), 486–507 (2015). doi:10.1177/0049124114543236

R.C. MacCallum, 2001 presidential address: working with imperfect models. Multivar. Behav. Res. **38**(1), 113–139 (2003). doi:10.1207/S15327906MBR3801_5

R.C. MacCallum, M.W. Browne, H.M. Sugawara, Power analysis and determination of sample size for covariance structure modeling. Psychol. Methods **1**(2), 130–149 (1996). doi:10.1037//1082-989X.1.2.130

H.W. Marsh, K.-T. Hau, Z. Wen, In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. Struct. Equ. Model. **11**(3), 320–341 (2004). doi:10.1207/s15328007sem1103_2

A.W. Meade, E.C. Johnson, P.W. Braddy, Power and sensitivity of alternative fit indices in tests of measurement invariance. J. Appl. Psychol. **93**(3), 568–592 (2008). doi:10.1037/0021-9010.93.3.568

J. Nevitt, G.R. Hancock, Evaluating small sample approaches for model test statistics in structural equation modeling. Multivar. Behav. Res. **39**(3), 439–478 (2004). doi:10.1207/S15327906MBR3903_3

D.L. Putnick, M.H. Bornstein, Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. Dev. Rev. **41**, 71–90 (2016). doi:10.1016/j.dr.2016.06.004

R Core Team, R: a language and environment for statistical computing (version 3.3.0). R Foundation for Statistical Computing (2016). Available via CRAN, https://www.R-project.org/

J.L. Rodgers, The bootstrap, the jackknife, and the randomization test: a sampling taxonomy. Multivar. Behav. Res. **34**(4), 441–456 (1999). doi:10.1207/S15327906MBR3404_2

Y. Rosseel, Lavaan: an R package for structural equation modeling. J. Stat. Softw. **48**(2), 1–36 (2012.) http://www.jstatsoft.org/v48/i02/

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

J.H. Steiger, J.C. Lind, Statistically-based tests for the number of common factors. Paper presented at the annual meeting of the Psychometric Society, Iowa City (1980)

semTools Contributors, semTools: useful tools for structural equation modeling (version 0.4–12) (2016). Available via CRAN. https://www.R-project.org/

R.J. Vandenberg, C.E. Lance, A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. Organ. Res. Methods **3**(1), 4–70 (2000). doi:10.1177/109442810031002