Permutation Randomization Methods for Testing Measurement Equivalence and Detecting Differential Item Functioning in Multiple-Group Confirmatory Factor Analysis

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

DOI
10.1037/met0000152

Publication date
2018

Document Version
Final published version

Published in
Psychological Methods

Psychological Methods
2018, Vol. 23, No. 4, 708–728

1082-989X/18/$12.00 http://dx.doi.org/10.1037/met0000152

Permutation Randomization Methods for Testing Measurement Equivalence and Detecting Differential Item Functioning in Multiple-Group Confirmatory Factor Analysis

Terrence D. Jorgensen
University of Amsterdam

Benjamin A. Kite and Po-Yi Chen
University of Kansas

Stephen D. Short
College of Charleston

Abstract

In multigroup factor analysis, different levels of measurement invariance are accepted as tenable when researchers observe a nonsignificant $(\Delta)\chi^2$ test after imposing certain equality constraints across groups. Large samples yield high power to detect negligible misspecifications, so many researchers prefer alternative fit indices (AFIs). Fixed cutoffs have been proposed for evaluating the effect of invariance constraints on change in AFIs (e.g., Chen, 2007; Cheung & Rensvold, 2002; Meade, Johnson, & Braddy, 2008). We demonstrate that all of these cutoffs have inconsistent Type I error rates. As a solution, we propose replacing χ^2 and fixed AFI cutoffs with permutation tests. Randomly permuting group assignment results in average between-groups differences of zero, so iterative permutation yields an empirical distribution of any fit measure under the null hypothesis of invariance across groups. Our simulations show that the permutation test of configural invariance controls Type I error rates better than χ^2 or AFIs when the model contains parsimony error (i.e., negligible misspecification) but the factor structure is equivalent across groups (i.e., the null hypothesis is true). For testing metric and scalar invariance, $\Delta\chi^2$ and permutation yield similar power and nominal Type I error rates, whereas ΔAFIs yield inflated errors in smaller samples. Permuting the maximum modification index among equality constraints control familywise Type I error rates when testing multiple indicators for lack of invariance, but provide similar power as using a Bonferroni adjustment. An applied example and syntax for software are provided.

Translational Abstract

Researchers measuring psychological constructs may first want to examine whether the items used to measure these constructs are interpreted equivalently in different contexts. Multiple-group factor analysis is a popular tool for testing measurement equivalence/invariance across populations, but methods for detecting item differences, such as observing a nonsignificant $(\Delta)\chi^2$ test after imposing equality constraints, may detect negligible differences in large samples. Researchers have instead suggested the use of alternative fit indices (AFIs) and have proposed fixed cutoffs for change in AFIs when evaluating invariance (e.g., Chen, 2007; Cheung & Rensvold, 2002; Meade, Johnson, & Braddy, 2008). Alternatively, we propose replacing χ^2 and fixed AFI cutoffs with permutation tests. Our simulations show that permutation tests yielded well controlled Type I error rates even when the model does not fit perfectly, providing the only valid test of configural invariance across groups of which we are currently aware. In addition, regardless of which AFI is preferred for tests of metric (i.e., item factor loadings) or scalar (i.e., item intercepts) equivalence, permutation tests provide well controlled Type I error rates, with power to detect true differences that is comparable with $\Delta\chi^2$. Conversely, we do not recommend that researchers use fixed cutoffs for ΔAFIs because our results suggest these cutoffs lead to inflated Type I error rates at smaller sample sizes. To encourage applications of the permutation procedure for testing measurement equivalence, we provide a complete real-data example, including software syntax for analysis and an interpretation of the results.

Keywords: measurement equivalence, configural invariance, differential item functioning, permutation, multiple group confirmatory factor analysis

Supplemental materials: http://dx.doi.org/10.1037/met0000152.supp

This article was published Online First November 27, 2017.

Terrence D. Jorgensen, Department of Child Education and Development, University of Amsterdam; Benjamin A. Kite and Po-Yi Chen, Department of Psychology, University of Kansas; Stephen D. Short, Department of Psychology, College of Charleston.

These results were presented as a paper presentation July 2016 at the 81st annual International Meeting of the Psychometric Society (IMPS), in Asheville, NC, and a portion of the results (configural invariance only) were published in the proceedings for that conference (see References).

The R syntax used to generate data for the Monte Carlo simulations is available from the first author upon request. We thank SURFsara (www.surfsara.nl) for the support in using the Lisa Compute Cluster to conduct our Monte Carlo simulations.

Correspondence concerning this article should be addressed to Terrence D. Jorgensen, Department of Child Education and Development, University of Amsterdam, Postbus 15776, 1001NG Amsterdam, the Netherlands. E-mail: t.d.jorgensen@uva.nl

Measurement equivalence/invariance (ME/I) is a key concept in psychological measurement (Brown, 2015) concerning whether indicators used to measure latent constructs have the same meaning across different contexts, such as different occasions or populations (Meade & Lautenschlager, 2004). The assumption of ME/I must be satisfied before groups or occasions can be compared on common factors. Previous studies have shown that failure to meet this assumption can result in selection bias or false detection of group differences (Chen, 2008; Millsap & Kwok, 2004). Given the importance of ME/I, various methods have been developed to examine it. In the current study, we focus on the multigroup confirmatory factor analysis (CFA) framework, which is one of the most common frameworks used to test ME/I across groups (Vandenberg & Lance, 2000).

This article is organized as follows. We begin by describing the current recommended best practices for testing ME/I, as well as discussing their limitations. We then introduce permutation randomization tests before proposing a permutation framework for testing ME/I across groups. We present Monte Carlo simulation studies to compare the power and Type I error rates of the permutation method to other methods, which we organize into three sections: testing equivalence of model form, testing equivalence of sets of model parameters, and detecting differences in individual indicator parameters. We illustrate the permutation procedure in a real-data application, and we conclude with recommendations for applied researchers and for further development of permutation methods for tests of ME/I (or lack thereof).

Testing Full and Partial Measurement Invariance

In the context of CFA, researchers commonly test three or four ME/I models in a sequence: configural, metric, scalar, and strict 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 χ^2 test statistic (Byrne, Shavelson, & Muthén, 1989; Meredith, 1993) is used to judge whether the configural invariance model is supported. If the test is not significant at the specified α level, the analyst can conclude that the configural model fits well and proceed to test metric (or "weak") invariance by constraining factor loadings to equality across groups. Metric invariance assumptions are deemed tenable if the change in fit ($\Delta\chi^2$ between configural and metric models) is not significant, in which case valid comparisons of latent variances and covariances can be made (and equivalence can be tested) across groups. Similar criteria and procedures can be used to test scalar (or "strong") invariance by constraining intercepts or thresholds to equality across groups, or to test strict invariance by constraining residual variances to equality across groups. Scalar invariance is required for valid comparisons of latent means to be made. Latent-parameter comparisons do not require strict invariance, but researchers might be interested in that test if they are interested in whether the reliability of a scale is equivalent across groups.

The $\Delta\chi^2$ tests for metric, scalar, and strict ME/I are "omnibus" in the sense that they test several parameter constraints simultaneously. Like using ANOVA for multiple group-mean comparisons, rejecting the null hypothesis (H_0) for the omnibus test typically necessitates multiple follow-up hypotheses to be tested (e.g., all possible pairwise comparisons, complex contrasts). If the omnibus H_0 of ME/I is rejected, it is still possible to make valid comparisons of latent parameters between groups, as long as some of the measurement parameters can be constrained to equality across groups. Constraining some, but not all, measurement parameters is often called *partial invariance* (Byrne et al., 1989; Steenkamp & Baumgartner, 1998). To establish partial invariance, it is necessary to identify which parameters substantially differ across groups. We borrow the term *differential item* (or *indicator*) *functioning*[1] (DIF) from the item-response theory (IRT) literature to refer to differences in measurement parameters across groups.

The CFA literature provides some guidance on the detection of DIF, most of which involves investigation of modification indices and expected parameter changes (Byrne et al., 1989; Steenkamp & Baumgartner, 1998). Other proposed methods to detect DIF include fitting a series of models (e.g., releasing equality constraints for one indicator at a time and testing $\Delta\chi^2$ compared with the fully constrained model) and calculating a test statistic for the difference between groups' factor loadings (Meade & Bauer, 2007). Modification indices approximate the change in the χ^2 statistic if a constraint is released so they accomplish the same goal as the model-comparison approach without having to fit several models. According to Byrne, Shavelson, and Muthén (1989), researchers can release the across-groups equality constraint which has the largest expected decrease in the χ^2 statistic, as estimated by modification indices. For example, if the H_0 of metric invariance is rejected by the data, then modification indices for constrained factor loadings should be investigated to indicate which loadings should be freely estimated across groups. A partial metric invariance model can be identified by releasing the fewest constraints that result in similar fit as the model with no constraints on loadings (i.e., nonsignificant $\Delta\chi^2$ compared with the configural model).

Limitations of Current Best Practices

The χ^2 statistic confounds two sources of model misfit (Cudeck & 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, χ^2 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

[1] Although "item" typically refers to a discretely measured test item, DIF has also been used to refer to measurement-parameter differences in the context of continuous indicators in CFA (e.g., Gonzalez-Roma, Tomas, Ferreres, & Hernandez, 2005; Kim & Willson, 2014), and Kline (2011, p. 253) referred to differentially functioning "indicators." Other terms have been used for the same phenomenon, such as *measurement bias* (Jak, Oort, & Dolan, 2010; Millsap, 2011, p. 47). Byrne et al. (1989) referred to differences in measurement parameters as *noninvariance*, whereas Meredith (1993) avoided using a double-negative prefix by introducing the term *structural bias*. *Response shift* in the health-related quality-of-life literature refers to changes in measurement properties over time; this would require extending the current method to longitudinal CFA.

even probable) that an analysis model corresponds only approximately to the groups' population models, yet the analysis model is equally (in)appropriate for each group (Jorgensen, Kite, Chen, & Short, 2017). 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 χ^2 does not truly provide a test of configural invariance.

Consider an example that helps illustrate this distinction. The population factor loadings in Table 1 represent a four-factor CFA model with approximate simple structure (i.e., the first four indicators have high loadings only in the first factor, the second four indicators have high loadings only on the second factor, etc.). However, all 16 indicators have small nonzero factor loadings on the remaining factors. If we fit a four-factor CFA without any cross-loadings, the H_0 of perfect fit would be false, and the power to detect the discrepancy from simple structure would increase with sample size. But suppose we measure these indicators in samples of both men and women, and that simple structure is approximately (but not exactly) correct in both populations. The H_0 of perfect correspondence between population and analysis models would still be false (arguably, to a degree of little practical consequence), but the H_0 of configural invariance would be true because the analysis model approximates each population's data-generating process equally well. A valid test of configural invariance should therefore fail to reject the H_0 of group equivalence even if the H_0 of perfect overall model fit were rejected. In such a situation, we would still need to assess whether the model fits adequately well enough to serve as a useful approximation of the real data-generating process (i.e., whether our model makes predictions that correspond closely with observations). If model modification is deemed necessary, ruling out group differences (i.e., failing to find evidence against configural invariance) could simplify the process, for example, by freeing parameters simultaneously in all groups instead of conducting independent specification searches within each group's model.

Table 1
Population Factor Loadings (Λ Matrix)

Indicator	Factor 1	Factor 2	Factor 3	Factor 4
1	.68 (.54)	−.03	−.02	−.11
2	.76 (.62)	.02	−.03	−.03
3	.74 (.60)	−.04	−.03	.00
4	.75 (.61)	.00	−.01	.08
5	.04	.76 (.61)	.07	.00
6	−.06	**.56** (**.41**)	−.03	.04
7	−.08	**.75** (**.60**)	.07	.06
8	−.02	.72 (.57)	.05	−.03
9	.07	−.01	.80 (.65)	.00
10	−.01	−.03	**.58** (**.43**)	−.02
11	−.04	.06	**.80** (**.65**)	.03
12	.04	.00	.39 (.24)	.05
13	−.02	−.02	−.01	.65 (.51)
14	.00	−.13	−.03	.67 (.53)
15	.00	.03	−.01	.59 (.45)
16	.00	.03	.02	.67 (.53)

Note. For conditions with eight indicators per factor, λs in parentheses were used as population parameters, and λs for Indicators 17–32 were identical to λs for Indicators 1–16. Bold indicates indicators with DIF when investigating power for metric and scalar invariance. Cells with only one value (near zero) are minor discrepancies from simple structure (approximation error).

Large sample sizes make $(\Delta)\chi^2$ sensitive even to minute differences in model form or measurement parameters, which have little or no practical consequence on latent parameter estimates (Brannick, 1995; Meade & Bauer, 2007). Many researchers would prefer to use an alternative fit index (AFI) to assess the approximate fit of the configural model or to judge the similarity in approximate fit between nested models. AFIs could be thought of as measures of effect size or "practical" significance—indicating the degree to which a model's predicted values deviate from observed values—that accompany the test of statistical significance provided by $(\Delta)\chi^2$. Researchers often (Putnick & Bornstein, 2016) find it justifiable to use a theoretically derived model whose predictions approximate observations quite closely, even if the small discrepancies are statistically significant, and multiple AFIs could be used to characterize this approximation in different ways. Putnick and Bornstein's (2016) review of 126 articles over a 1-year period indicates that only 17% of ME/I tests are decided by $(\Delta)\chi^2$ alone, whereas 46% also involve at least one AFI, and 34% are decided using AFIs alone.

Dozens of AFIs have been proposed since the early 1970s, but we will focus only on the few that have been recommended for use in assessing ME/I: the comparative fit index (CFI; Bentler, 1990), McDonald's (1989) centrality index (Mc), the root mean square error of approximation (RMSEA; Steiger & Lind, 1980), and the standardized root-mean-square residual (SRMR; Bollen, 1989, p. 258). CFI was reported for 73.2% of ME/I tests, making it the most popular AFI in this context (Putnick & Bornstein, 2016). AFIs reflect 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 $(\Delta)\chi^2$. In the example above (involving the parameters in Table 1), suppose the χ^2 test was significant and that the CFI and RMSEA did not meet the conventional criteria for acceptable approximate fit. The H_0 of configural invariance is still true, and failing to reject the H_0 of group equivalence would imply that model modification would be required for all groups, not just one group. AFIs have additional limitations, some of which are discussed below.

Most AFIs do not have known sampling distributions,[2] so evaluating the fit of a configural model involves some subjective decisions (e.g., which fit indices to use, what values indicate acceptable fit), often using arbitrary rules of thumb (e.g., Mc > .90 or SRMR < .08; Hu & Bentler, 1999). Sometimes there are conflicting recommendations based on different criteria. For example, Bentler and Bonett (1980) suggested[3] CFI > .90 indicates good fit, yet Hu and Bentler (1999) recommended CFI > .95 as a stricter criterion. Browne and Cudeck (1992) suggested RMSEA < .05 indicates close fit, RMSEA < .08 indicates reasonable fit, and RMSEA > .10 indices poor fit (RMSEA between .08–.10 indicates mediocre fit; MacCallum, Browne, & Sugawara, 1996); yet Hu and Bentler (1999) recommended RMSEA < .06 as a stricter criterion. According to an August, 2016 Google Scholar search, Hu and Bentler's (1999) criteria seem to be more widely applied

[2] A notable exception is RMSEA. See an excellent discussion by Kenny, Kaniskan, and McCoach (2015).

[3] CFI itself was not proposed until 1990, but Bentler and Bonett's (1980) recommendation applied to incremental fit indices in general, of which CFI is the most popularly applied in the context of ME/I (Putnick & Bornstein, 2016).

(34,425 citations) than Bentler and Bonett's (1980) (13,632 citations) or Browne and Cudeck's (1992) (2,774 citations).

Cutoff criteria for change (Δ) in AFIs to compare nested models (e.g., metric or scalar equivalence) have also been proposed. Cheung and Rensvold (2002) recommended $\Delta CFI > -.01$ or $\Delta Mc > -.02$ based only on controlling Type I error rates. Meade, Johnson, and Braddy (2008) recommended $\Delta CFI > -.002$ to yield greater power, and provided variable cutoffs for ΔMc based on the number of indicators and factors. Under different simulation conditions, Chen (2007) recommended a general $\Delta CFI > -.005$ for metric and scalar equivalence, but recommended ranges of ΔSRMR between .005 and .03 or ΔRMSEA between .005 and .015 depending on sample size, sample size ratio, and whether metric or scalar equivalence was tested. Putnick and Bornstein (2016) did not report the frequency with which these different criteria were used, but an August, 2016 Google Scholar search suggested that Cheung and Rensvold's (2002) criteria (4,594 citations) have been applied more often than Chen's (2007) (1,144 citations) or Meade et al.'s (2008) (343 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, magnitude of population parameters, number and magnitude of omitted variables and parameters from the analysis model, data distributions, and so forth (Cheung & Lau, 2012; Pornprasertmanit, Wu, & Little, 2013). Even when Chen (2007) or Meade et al. (2008) provided variable cutoffs to accommodate some of these factors, they could only provide recommendations for the conditions they investigated, which might differ from an applied researcher's situation (e.g., like Cheung & Rensvold, 2002, they simulated only two groups). And although they recommended constant (yet conflicting) ΔCFI cutoffs, these cutoffs fail to control Type I errors because even when the expected values ΔAFIs are not sensitive to *N*, the larger sampling variance at smaller *N* yields a greater probability of rejecting a true H_0 (see Marsh, Hau, & Wen, 2004). Fixed cutoffs can also lead to the apparent paradox that larger samples yield lower power, which occurs when the (Δ)AFI cutoff is more extreme than the population-level (Δ)AFI.[4]

Limitations of DIF Detection Method

A body of literature (e.g., French & Finch, 2008; MacCallum, Roznowski, & Necowitz, 1992) has demonstrated that data-driven rather than theory-driven use of modification indices leads to models that are overfit to nuances of sample data, rather than to models that generalize to new samples from the same population. Using modification indices to detect DIF in the absence of theoretical guidance, researchers must screen indicators one-by-one for DIF, rather than freeing all indicators flagged for DIF in the full metric invariance model (Yoon & Kim, 2014). Because this procedure might involve multiple steps, preventing the inflation of Type I error rates requires a Bonferroni-like correction to the α level (French & Finch, 2008). This has implications for loss of power: The number of indicator parameters to test increases with the number of indicators,[5] and the number of constraints per parameter increases with the number of groups. If one were to divide the overall α level by the total number of modification indices to be inspected, preventing inflation of Type I errors would yield unacceptably high Type II error rates.[6]

Analogously, a Bonferroni correction is not the preferred method of controlling Type I errors following a significant ANOVA when there are several groups. Rather, a less conservative approach (e.g., Tukey or Scheffé tests) would still control Type I error rates but would yield greater power to detect group differences. A similar method would be desirable in the context of detecting DIF, and in the next section we show how permutation provides a method with these desirable characteristics. We begin with a brief introduction to permutation tests in general.

Permutation Randomization

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 method that can be used to create an empirical approximation of a sampling distribution is the permutation randomization test. A current disadvantage of permutation tests (and of resampling methods in general) is the increased computing time needed to form empirical distributions by resampling the data hundreds or thousands of times and applying the same statistical procedure (e.g., fitting the same models) to each of the permuted data sets. But the advantage of permutation tests is their flexibility. If a method of resampling the data can be conceived such that a H_0 is known to be true (in the permutation distribution), then reference distributions can be empirically approximated for statistics whose sampling distributions are unknown or intractable.

A simple example of a permutation test is to compare two group means. An independent-samples *t* test can be conducted under assumptions of independence, normality, and homoscedasticity, using student's *t* distribution with the appropriate degrees of freedom to calculate the probability (i.e., the *p* value) of a *t* statistic at least as extreme as the observed one, on the condition H_0 is true. If the data are not approximately normally distributed, the theoretical *t* distribution is no longer an accurate representation of how the *t* statistic truly varies across samples from the same population,

[4] Population-level (Δ)AFIs can be obtained by fitting the analysis model(s) to the population moments, or can be estimated from the average (Δ)AFI across Monte Carlo samples.

[5] Methods could be used to reduce the number of follow-up tests, rather than investigating all possible items. For example, French and Finch (2008) and Woods (2009) propose ways of identifying sets of invariant indicators, and Millsap and Olivera-Aguilar (2012) proposed measures of effect size that could be used to narrow down indicators with substantial DIF.

[6] Cheung and Lau (2012) described a method to define differences between groups' measurement parameters as functions of other model parameters. Differences can then be tested using bias-corrected bootstrap confidence intervals, although the confidence level also requires a Bonferroni-like correction. Asparouhov and Muthén (2014) developed an alignment technique similar to rotation of an EFA solution, but freely estimated factor loadings in the configural model are transformed using an algorithm that assumes most indicators are invariant and few indicators (but more than zero) have substantial DIF. The alignment method also requires using a very conservative α level (e.g., $\alpha = .001$), although they do not recommend employing a Bonferroni correction because large numbers of pairwise comparisons render tests too conservative.

leading to incorrect p values and thus Type I error rates that are inflated or deflated relative to the specified α level. Although in practice the t test is robust to moderate departures from normality, the true shape of the statistic's sampling distribution under H_0 is generally unknown for nonnormal populations. The sampling distribution can be empirically approximated using resampling methods such as Monte Carlo simulation, permutation, or bootstrapping (Rodgers, 1999).

The logic of the permutation test is related to the use of random assignment in experimental designs. Random assignment of subjects to two groups will average out any between-groups differences, so that on average, group mean-differences would be zero, resulting in two 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[7] or experimental), any existing between-groups differences would be zero, on average.

To accomplish this, 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.

The value of permutation methods lies in their flexibility. With multiple groups, ANOVA's F statistic can be permuted to test the omnibus H_0, and permutation can also be used to control the Type I error rate for post hoc pairwise comparisons (see Higgins, 2004). For example, if there are four groups, then there are six possible pairwise comparisons to test whether the omnibus H_0 is rejected. Assuming H_0 is true (i.e., a Type I error was made), then maintaining a nominal Type I error rate in the follow-up tests would entail choosing a critical value such that only the largest among the six absolute differences would be rejected in $(\alpha \times 100)\%$ of samples. This is the motivation behind Tukey's honest significant difference (HSD) post hoc procedure:[8] Instead of comparing a t statistic for each mean difference to a t distribution, each t statistic is compared with a critical value from the studentized range distribution, which takes into account how many t statistics are being tested simultaneously. Permutation methods can approximate a similar distribution by calculating all possible pairwise mean-differences at each permutation, saving only the largest in absolute value. Then p values can be calculated for each observed mean-difference by calculating the proportion of the permutation distribution that exceeds the absolute value of the observed mean-difference.

Permutation methods have recently been developed in the context of factor analysis and structural equation modeling (SEM). These include testing the contribution of each variable to a solution in principle components analysis (Linting, van Os, & Meulman, 2011), detecting switched latent-class labels in simulation studies involving finite mixture models (Tueller, Drotar, & Lubke, 2011), and testing residual correlations among adjacent questionnaire items (Hildreth, Genschel, Lorenz, & Lesser, 2013). We propose a permutation method for testing ME/I and detecting DIF, which are simple extensions of the examples above.

Permutation tests of ME/I

Randomly permuting group assignment yields resampled data for which the H_0 of group equivalence is true. The steps to test configural ME/I are similar to the permutation test of means described above:

1. Fit the hypothesized multiple-group model(s) 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 the original data to the nth value from the new group vector $G_{\text{perm}(i)}$. 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).
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 with the vector of permuted fit measures.

Step 6 can be accomplished in either of two ways, yielding the same decision about H_0:

- Calculate the proportion of the vector of permuted fit measures that is more extreme (i.e., indicates worse fit or a greater decrement in fit) than the observed fit measure. This is a one-tailed p value[9] 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, SRMR, or

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

[8] Also referred to as Tukey's wholly significant difference (WSD) procedure.

[9] An exact p value could be calculated from all possible permutations of group assignment; however, this would become computationally intractable as the sample size or number of groups increases. For example, with only 50 people in each of two groups, the number of possible permutations would be on the order of 3.07×10^{93} (not quite a googol, but pretty big). However, the p value calculated from a large random sample of all possible permutations is a good estimate of the exact p value.

RMSEA; or sort in descending order for goodness of fit indices like CFI or Mc. Use the [100 × (1 - α)]th percentile as a critical value, and reject H_0 if the observed fit measure is more extreme than the critical value.

Testing metric, scalar, or strict ME/I entails the same steps, but the nested (restricted and unrestricted) models are both fit at Steps 1 and 4, and differences in fit measures are saved (e.g., $\Delta\chi^2 = \chi^2_{\text{restricted}} - \chi^2_{\text{unrestricted}}$, or $\Delta\text{AFI} = \text{AFI}_{\text{restricted}} - \text{AFI}_{\text{unrestricted}}$).

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. Randomly reassigning observations to groups results in samples whose patterns in their covariance matrices (i.e., which variances are larger than others, which variables are most or least strongly correlated, the direction of correlations) are consistent across groups, on average. Irrespective of whether the hypothesized model can reproduce those observed patterns perfectly, approximately well, or poorly, the model is expected to fit equally well (or equally poorly) in all groups in a permuted data set. Thus, the permutation distribution of a fit measure (χ^2 or any AFI) reflects the fit of the model to all groups, on the assumption that all group structures are equivalent. If the H_0 of equivalent structures across groups is true, then the observed fit measure would only rarely (defined by the specified α level) be inconsistent with the values in the permutation distribution. However, if the groups' structures substantially differ, then the observed fit measure would be inconsistent with the values in the permutation distribution of that fit measure. Naturally, the power to detect the inconsistency between group structures would depend on the degree of inconsistency, the sample size, and the test criterion (defined by α).

Furthermore, permutation provides empirical sampling distributions of (Δ)AFIs, which generally have unknown sampling distributions. Researchers using permutation methods would not need to rely on fixed cutoff criteria proposed from studies whose simulated conditions might not closely resemble their own data and model(s), such as $\Delta\text{CFI} < -.01$ (Cheung & Rensvold, 2002), $\Delta\text{CFI} < -.005$ (Chen, 2007), or $\Delta\text{CFI} < -.002$ (Meade et al., 2008). 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(s) under consideration. Because model fit is unaffected by the method of identifying the location and scale of the latent construct(s), results of the permutation method are independent of whether a researcher chooses to use a standardized metric, a marker/reference variable, or effects-coding (Little, Slegers, & Card, 2006).

Permutation Tests of DIF

Assuming a well-fitting model and no evidence against the H_0 of configural invariance, researchers may find evidence against the H_0 of more restrictive levels of ME/I in subsequent steps, which would be evidence that not all measurement parameters are equivalent across groups. For example, if adding the constraint of equal factor loadings across groups causes fit to deteriorate enough for $\Delta\chi^2$ to be significant, then that would constitute evidence against the H_0 of metric ME/I. Researchers wishing to establish partial ME/I must choose which indicator parameter to freely estimate across groups before permuting groups again to compare, for example, a partial metric invariance model to the configural model. Because this typically entails looking at multiple modification indices from the restricted model fit in Step 1, the same indices can be saved for permuted data in Step 4, resulting in an empirical distribution of each modification index under H_0. However, simultaneously testing multiple modification indices increases the probability of committing a Type I error (falsely concluding that DIF is present more often than a predetermined acceptable error rate: α). Even under a false omnibus H_0, the error rate would still be inflated for follow-up tests whenever there is more than one true H_0 about individual parameters.[10]

Although a Bonferroni-adjusted α level can be used to control familywise Type I errors (French & Finch, 2008; Jak et al., 2010), a method like Tukey's HSD is likely to be more powerful when the number of indicators or groups is large. Using the same logic behind Tukey's studentized range distribution, we desire a method that behaves as follows. When the omnibus H_0 is true, any modification index judged to be significant would constitute a Type I error. In some samples, multiple modification indices (i.e., multiple potential Type I errors) could be significant. In principal, in a sample with at least one potential Type I error, the largest modification index must be among them, whereas in a sample with only one potential Type I error, the largest modification index must be the one that results in that error. This implies that if we were to use a test criterion that only committed a Type I error in (100 × α)% of samples (by detecting the largest modification index to be significant under a true H_0), then any smaller modification indices would be detected as significant in no more (and probably fewer) than (100 × α)% of samples. Such a test criterion would therefore keep the familywise Type I error rate at the nominal α level.

Permutation again provides a simple framework to accomplish this. At Step 4, instead of saving all modification indices of interest from a permuted data set, save only the largest among that set (e.g., if the omnibus H_0 of metric invariance is rejected, save the largest modification index involving equality-constrained factor loadings). The resulting distribution of the largest modification index observed under H_0 provides a p value or critical value that implicitly adjusts for the number of indices tested, as described in the previous paragraph.

Monte Carlo Simulations

To evaluate the permutation methods proposed in the previous section, we present results from a series of small-scale simulation studies. They are organized in an order that mimics the procedure of testing ME/I in practice. The first two simulation studies involved testing configural invariance when H_0 is true (to ascertain Type I error rates) and false (to estimate power). The next simulations ascertained rejection rates (i.e., Type I error rates when H_0 is true, power when H_0 is false) of omnibus tests of metric and scalar invariance. In cases when full metric or scalar invariance is deemed untenable, partial invariance could be established by testing individual indicators for DIF. Thus, the final simulation studies investigated rejection rates for modification indices associated with across-group constraints in factor loadings or intercepts.

[10] This is analogous to the failure of Fisher's LSD post hoc test to control Type I error rates following a significant omnibus F test in ANOVA.

We designed each simulation study to compare H_0 rejection rates between permutation methods and currently recommended practices under a variety of conditions. To demonstrate whether permutation methods could supplant currently recommended cutoffs for (Δ)AFIs, we chose conditions that partially replicated those of Meade et al. (2008).[11] We present all of our design factors here, at least two of which are manipulated in each small-scale study.

Monte Carlo Design Factors

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. The analysis models included CFA models with varying levels of ME/I constraints (configural, metric, and scalar), as well as an appropriate null model for calculating CFI (see Widaman & Thompson, 2003). Using lavaan's default settings, the scales of the latent factors were identified by fixing the first factor loading (which was always invariant across groups in each condition) to one, and the latent means were fixed to zero (except in the scalar model, in which Group 1's latent mean was fixed to zero, but Group 2's latent mean was freely estimated). Rather than using the default independence model (i.e., constrain all covariances to zero, but freely estimate all variances and means separately in each group), we constrained variances and means to equality across groups, so that the null model would be nested within all CFA models.[12]

Based on Meade et al. (2008), we varied sample size (N) in each of two groups across five levels: 100, 200, 400, 800, and 1,600 per group. We excluded Meade et al.'s asymptotic condition (N = 6,400 per group). We varied model complexity via number of factors (two or four) and number of indicators per factor (four or eight), using the same population values for factor loadings as Meade et al. (see Table 1) so that overall scale reliability was constant across conditions. Although the model has simple structure, Meade et al. sampled nonsalient cross-loadings (normal with $\mu = 0$, $\sigma = 0.05$) to include approximation discrepancy, the presence of which prevents a traditional χ^2 test or AFI from assessing configural invariance independently from overall model fit.[13] We fixed all intercepts to zero (except in DIF conditions, discussed next), factor means to zero, factor variances to one, factor correlations to 0.3, and residual variances to values that would set total indicator variances to one (i.e., standard normal variables) when DIF was absent.

Based on Meade et al. (2008), we varied lack of invariance (LOI) independently for configural, metric, and scalar models. Like Meade et al. we simulated metric or scalar LOI by manipulating the magnitude of DIF between 0 and 0.4 for both factor loadings and intercepts.[14] However, we used increments of 0.1 instead of 0.02, preferring to simulate more replications in fewer conditions. In contrast to Meade et al.'s 500 replications per condition, we simulated 2,000 replications to minimize Monte Carlo sampling error of estimated rejection rates and critical values. In each DIF condition, the magnitude of DIF was subtracted from Group 2's loading or intercept of the first bolded indicator per factor in Table 1, and the same magnitude of DIF was added to Group 2's second bolded indicator per factor in Table 1; thus, the total number of differentially functioning indicators varied with the number of factors, but the proportion of indicators with DIF was constant (25%). Residual variances remained constant across all conditions, so total variances of differentially functioning indicators could differ from 1.0 when DIF >0.

Whereas Meade et al. (2008) simulated configural LOI by adding additional factors to Group 2'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 2'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 levels of configural LOI were arbitrary, but they only needed to serve as a basis for comparing the power of different methods to detect the same lack of correspondence between the groups' population models.

In all conditions, $I = 200$ permutations were used to calculate p values associated with (Δ)AFIs, as well as their critical values at $\alpha = .05$. Although applied researchers should use at least 1,000 permutations to reduce Monte Carlo sampling error of an estimated p value, we were concerned only with rejection rates across 2,000 replications, not with the precision of a single replication's approximate p value. We conducted a preliminary study (not presented here) to verify that rejection rates were similar using p values calculated from 200 permutations and 10,000 permutations. We now present results of each small-scale simulation, preceded by a description of the conditions of that study and the research questions it was designed to answer.

Testing Configural Invariance

Our first simulation study was designed to demonstrate how often the traditional χ^2 (or AFI rule of thumb) would reject a true H_0 of configural invariance under various conditions (see also Jorgensen et al., 2017). Because the population models included minor approximation discrepancy in the form of near-zero cross-loadings, the configural model did not fit perfectly even to the population data; thus, a traditional tests of configural invariance would confound overall model fit with group equivalence. We fit the configural model to population moments of both groups in each condition, and we provide population-level fit measures in Table 2 to verify that the approximation error was minor. The expected power of χ^2 (calculated using the method described in

[11] Although there were other studies available to partially replicate, we did not choose Cheung and Rensvold's (2002) design because they investigated only Type I error rates. Although Chen (2007) also investigated power, she only manipulated the number of indicators, whereas Meade et al. (2008) manipulated numbers of both indicators and factors. Meade et al. also included approximation discrepancy (as did Cheung & Rensvold, 2002), whereas Chen (2007) did not, and this feature is required to demonstrate the effectiveness of the permutation method for testing configural invariance.

[12] In practice, it would only be necessary to constrain variances across groups if strict invariance were also tested, which we did not do in our simulations.

[13] When the analysis model perfectly corresponds to the population model, asymptotically nominal Type I error rates for the χ^2 are well documented. Furthermore, we consider perfect fit to be unrealistic in practice, so we do not simulate data under that condition.

[14] Meade et al. (2008) only varied LOI between 0 and 0.3 for intercepts.

Table 2
Fit Measures for Configural Model Fit to Population Data When H_0 Is True

Fit measure	*N* per group	2 Factors		4 Factors	
		4 Indicators (*df* = 38)	8 Indicators (*df* = 206)	4 Indicators (*df* = 196)	8 Indicators (*df* = 916)
χ^2 (power)	100	4.37 (13%)	6.34 (9%)	20.53 (26%)	29.64 (17%)
	200	8.75 (25%)	12.67 (16%)	41.05 (61%)	59.28 (39%)
	400	17.49 (54%)	25.34 (33%)	82.11 (97%)	118.57 (84%)
	800	34.98 (92%)	50.69 (74%)	164.21 (100%)	237.14 (100%)
	1,600	69.97 (100%)	101.38 (99%)	328.42 (100%)	474.27 (100%)
CFI	≤800	1	1	1	1
	1,600	.996	1	.991	1
Mc	≤800	≥1	≥1	≥1	≥1
	1,600	.995	≥1	.980	≥1
RMSEA	≤800	0	0	0	0
	1,600	.023	0	.021	0
SRMR	≤1,600	.023	.029	.020	.024

Note. CFI = comparative fit index; Mc = McDonald's (1989) centrality index; RMSEA = Root mean square error of approximation; SRMR = standardized root-mean-square residual. Power for χ^2 calculated using method described in Satorra and Saris (1985).

Satorra & Saris, 1985) varied widely (9%–100%) depending on *N* and model complexity, but all AFIs showed excellent approximate fit according to Hu and Bentler's (1999) criteria.

We used a 5 (*N*) × 2 (two or four factors) × 2 (four or eight indicators per factor) design, holding LOI constant at zero. We expected Type I error rates to be inflated beyond 5%, and for these rates to increase with sample size. We had no specific hypotheses using fixed cutoffs for AFIs, but because fixed cutoffs do not take sampling variability or model complexity into account, we expected results to vary across *N*s and model sizes. Because permutation only removes group differences, we expected nominal Type I error rates in all conditions for all fit measures, which would indicate a valid test of configural invariance that is independent of overall model fit.

Results. As expected, using the traditional χ^2 test of exact fit to test configural invariance 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 *N*s, rejection rates matched the expected power using the Satorra and Saris (1985)

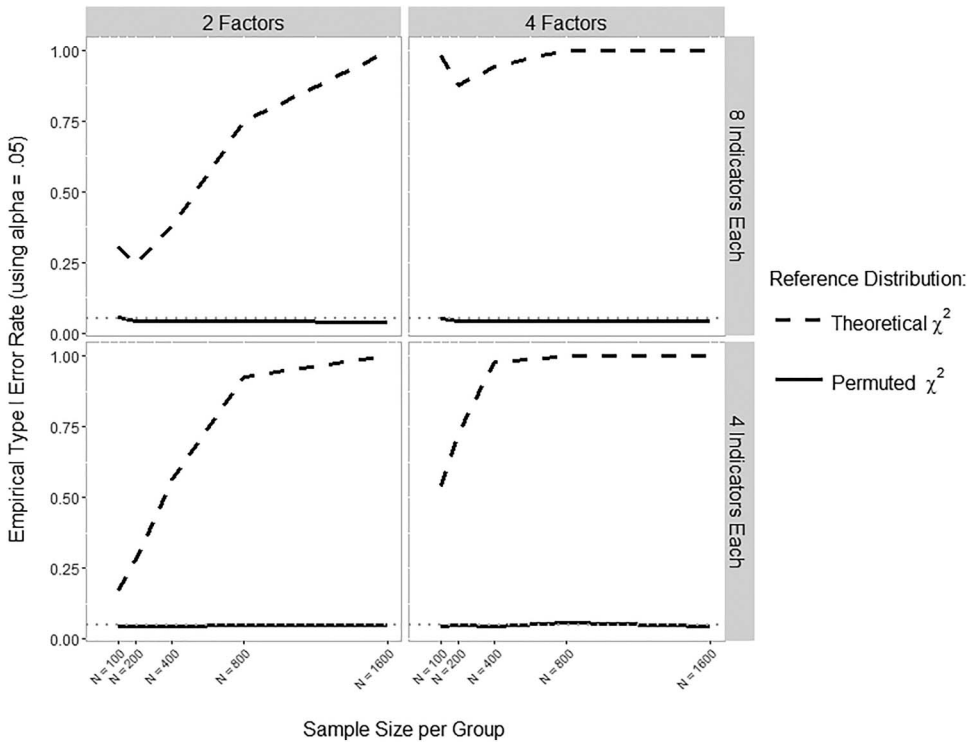

Figure 1. Type I error rates for χ^2 and permutation test of configural invariance, as well as expected power of the χ^2 using the Satorra and Saris (1985) method. The dotted gray line indicates the nominal error rate (5%).

method, but rejection rates were inflated at smaller *N*, especially in larger models, consistent with previous research demonstrating small-sample bias of χ^2 (Nevitt & Hancock, 2004). In contrast, the permutation method provided nominal Type I error rates across conditions.

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 2 shows that for the smallest model, fixed cutoffs for CFI almost always resulted in no Type I errors. With either additional factors or additional indicators, using Hu and Bentler's (1999) criterion inflated the Type I error rates when $N = 100$ per group. For the largest model, even Bentler and Bonett's (1980) less stringent criterion resulted in over 80% Type I errors at $N = 100$ per group. Similar results were found for other AFI guidelines[15] (RMSEA, SRMR, and Mc). In contrast, permuting CFI (or any AFI) maintained nominal Type I error rates across all conditions.

Power Analysis

Having established that permutation provides more consistent control of Type I error rates than χ^2 or AFI rules of thumb across a variety of conditions, we next investigated whether power to detect LOI using permutation is comparable to using χ^2 or AFI cutoffs. We used a 5 (*N*) × 4 (LOI) design, holding model complexity constant (four indicators for each of two factors, the condition in which fixed cutoffs for CFI showed ≤5% Type I errors). We expected permutation to have lower power than χ^2, 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 (displayed in Table 3) met guidelines for acceptable fit.

Results of power analysis. Figure 3 confirms our expectation that χ^2 had the highest power to detect LOI, particularly at the lowest level of LOI and the smallest *N*. But as Figure 1 shows, the greater power came at the expense of high Type I errors because χ^2 tests overall model fit rather than configural invariance alone. Hu and Bentler's (1999) more stringent criterion (CFI > .95) yielded power almost as high as χ^2, whereas Bentler and Bonett's (1980) less stringent criterion (CFI > .90) yielded lower power that decreased as *N* increased in conditions where only one or two salient population parameters differed between groups. This apparent contradiction occurs because the population CFI > .90 in those conditions (see Table 3), so less sampling variance resulted in fewer model rejections (Marsh et al., 2004). The online supplementary materials shows the same pattern of results for other AFIs (RMSEA, SRMR, Mc).

Permutation yielded inadequate power to detect the smallest amount of LOI (i.e., when a single parameter differs between populations), unless $N \geq 800$ per group. Adequate power to detect greater LOI was achieved at smaller *N*. The permuted χ^2 tended to have greater power than permuted CFI, but the discrepancy was small when *N* and LOI were large. Permuted RMSEA and Mc (depicted in online supplementary materials) had power similar to the permuted χ^2, but permuted SRMR had consistently low power.

Testing Invariance of Sets of Model Parameters

To investigate Type I error rates when testing metric invariance, we used a 5 (*N*) × 2 (two or four factors) × 2 (four or eight indicators per factor) design, holding DIF constant at zero. We expected nominal Type I error rates for the traditional and permuted $\Delta\chi^2$, and we also expected permutation of ΔAFIs to maintain nominal Type I error rates. Because fixed critical values for ΔAFIs do not take sampling error into account, we expected Type I error rates to vary across *N*s and model sizes. Our design and hypotheses for scalar invariance were the same as for metric, but results were so similar that we present scalar invariance results only in the online supplemental materials.

Results. Figure 4 confirms that Type I error rates were well controlled across conditions for the traditional (with some small-sample bias apparent in the larger models) and for permuted $\Delta\chi^2$. As expected, Figure 5 shows that permuting ΔCFI also maintained nominal Type I error rates, but fixed critical values did not provide consistent Type I error rates. Cheung and Rensvold's (2002) critical value (ΔCFI = −.01) yielded Type I error rates ≤5% only when $N \geq 200$ per group; Chen's (2007) critical value (ΔCFI = −.005) yielded Type I error rates ≤5% only when $N \geq 400$ per group; and Meade et al.'s (2008) critical value (ΔCFI = −.002) yielded Type I error rates ≤5% only when $N \geq 800$ per group. It is noteworthy that Cheung and Rensvold (2002) established critical value based only on controlling Type I error rates,[16] and theirs resulted in the fewest Type I errors. Chen (2007) and Meade et al. (2008) in contrast, developed guidelines with the intent of increasing power to detect LOI, and their guidelines resulted in highly inflated error rates unless *N* was large.

To demonstrate why permutation maintains nominal error rates across conditions, we plotted the observed 95th percentiles from the Monte Carlo distribution in each condition (dashed lines in Figure 6). Because the Monte Carlo distribution is an approximation of the true sampling distribution of ΔCFI, these critical values are the gold standard. For each replication, we saved the 95th percentile of the permutation distribution, and we calculated the average of these percentiles in each condition (solid lines in Figure 6). The close correspondence of the black lines in Figure 6 illustrate that permutation provides an unbiased estimate of the "true" critical value in a particular condition, yielding nominal Type I error rates. The intersection of each gray dotted line with the black lines indicates under which condition that fixed critical value might yield nominal Type I error rates in practice.

Cheung and Rensvold's (2002) ΔMc critical value (ΔMc = −.02) yielded similar Type I error rates in each condition as their ΔCFI critical value. Meade et al. (2008) provided variable ΔMc cutoffs across a range of models; their guidelines for models in the current study were approximately ΔMc > −.007, which yielded Type I error rates ≤5% only when $N \geq 400$ per group (see Figure 7). Permuting ΔMc yielded nominal error rates across conditions. Chen's (2007)

[15] Although not depicted here, supplementary figures are available in an online appendix.

[16] Cheung and Rensvold (2002) used 99th percentiles to choose cutoffs, corresponding to $\alpha = 1\%$, so even 5% is an inflated Type I error rate by that criterion.

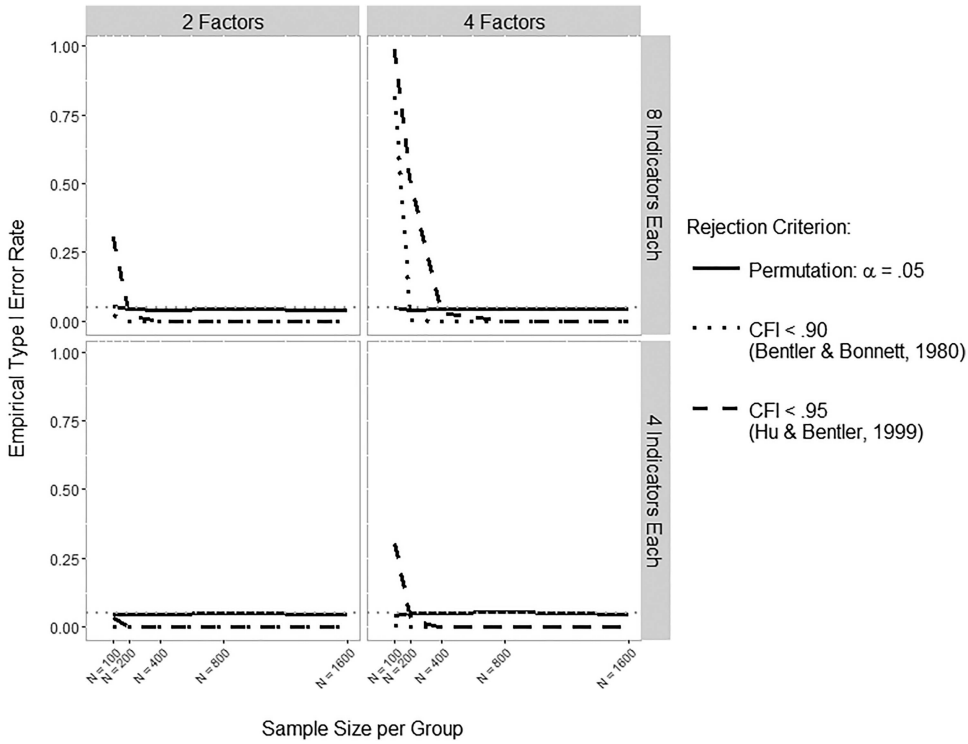

Figure 2. Type I error rates for CFI using fixed and permutation-based critical values. The dotted gray line indicates the nominal error rate (5%).

guidelines for ΔSRMR and ΔRMSEA yielded Type I error rates below 5% in all conditions (shown in online supplemental materials), with the exception that ΔRMSEA yielded 10% Type I errors when $N = 100$ per group in the two-factor, four-indicator condition. Again, permuting ΔSRMR and ΔRMSEA controlled Type I error rates well in all conditions.

Power Analysis

Having established that permutation provides more consistent control of Type I error rates than $\Delta\chi^2$ or fixed critical values for ΔAFIs across a variety of conditions, we investigated power to detect DIF. We used a 5 (N) × 4 (DIF) design, holding model complexity constant (four indicators for each of two factors, the condition in which fixed cutoffs for ΔCFI and ΔMc showed the least inflation of Type I errors). We expected $\Delta\chi^2$ to have similar power as the permuted $\Delta\chi^2$ and permuted ΔAFIs, given their similar rejection rates when H_0 was true. We expected fixed ΔAFI cutoffs to have low power when the population ΔAFI met criterion for good fit (see Table 4), but to have higher power than permuted ΔAFIs in conditions where Type I error rates were high, such as when N was small and Chen's (2007) and Meade et al.'s (2008) criteria were used. As with the investigation of Type I error rates, we used the same design and held the same hypotheses for scalar invariance as for metric, but we present only the result for metric invariance because scalar results are so similar. Results for scalar invariance are available in the online supplemental materials.

Results of Power Analysis

Figure 8 shows that power is nearly identical for the traditional and permuted $\Delta\chi^2$, as well as for any permuted ΔAFI. Power to detect negligible DIF (0.1) is low when $N = 100$ per group, but approaches 100% when $N = 1{,}600$ per group. For small DIF (0.2), the power ≥80% when $N \geq 400$ per group. Power to detect moderate (0.3) or large DIF (0.4) exceeds 80% when $N \geq 200$ or 100 per group, respectively.

Using fixed cutoffs for ΔAFIs, on the other hand, only provides a power advantage when sample sizes are small, which is when those cutoffs also yield inflated Type I error rates. Figure 9 shows that ΔCFI's power does not increase appreciably with N, as one would expect from a test based on a sampling distribution. As expected, Meade et al.'s (2008) criterion provides higher power to detect negligible DIF than permutation when $N \leq 400$ per group, as does Chen's (2007) criterion when $N = 100$ per group. However, the power of fixed cutoffs increases very little with N, so permutation provides greater power than fixed cutoffs with larger N. For negligible DIF, power actually decreases with N when using Cheung and Rensvold's (2002) or Chen's (2007) criteria; Cheung and Rensvold's (2002) criterion yields decreasing power even when DIF = 0.2. For moderate and large DIF, power curves are similar using permutation or fixed critical values.

We provide figures for additional ΔAFIs in the online supplemental materials, which show similar patterns when comparing the power curves of permuted ΔMc to those of Cheung and Rens-

Table 3
Fit Measures for Configural Model Fit to Population Data When H_0 Is False

Fit measure	*N* per group	Lack of invariance (additional parameters in Group-2 model) 1	2	3	4
$\chi^2_{df=38}$	100	40.50	50.11	70.30	80.93[a]
	200	80.99	100.21	140.59	161.86
	400	161.99	200.43	281.18	323.72
	800	323.98	400.86	562.37	647.43
	1,600	647.96	801.71	1124.74	1294.87
CFI	100	.995	.976	.941	.924
	200	.959	.942	.912	.896
	400	.943	.926	.899	.884
	800	.935	.919	.893	.878
	1,600	.932	.916	.890	.875
Mc	100	.994	.970	.922	.898
	200	.948	.925	.880	.857
	400	.925	.903	.859	.836
	800	.915	.893	.849	.827
	1,600	.909	.888	.844	.822
RMSEA	100	.026	.056	.092	.106
	200	.075	.090	.116	.128
	400	.090	.103	.126	.137
	800	.097	.109	.131	.142
	1,600	.100	.112	.134	.144
SRMR	≤1,600	.058	.057	.061	.059

Note. CFI = comparative fit index; Mc = McDonald's (1989) centrality index; RMSEA = Root mean square error of approximation; SRMR = standardized root-mean-square residual. Power for $\chi^2 > 99\%$, unless otherwise indicated.
[a] Power for $\chi^2 > 96\%$.

vold's (2002) and Meade et al.'s (2008) criteria for ΔMc, as well as between Chen's (2007) ΔRMSEA criteria and permuted ΔRMSEA. However, Chen's (2007) ΔSRMR criteria always yielded lower power than permuted ΔSRMR, even when $N = 100$ per group.

Testing Invariance of Individual Model Parameters

Upon rejecting the omnibus test of metric or scalar equivalence, a researcher's next step would be to test individual equality constraints. If releasing the across-group equality constraints for a minority of indicators leads to similar fit between the restricted and unrestricted model, then the researcher will have established partial invariance. Given that the omnibus $\Delta\chi^2$ and permutation tests yielded similar power to detect DIF in factor loadings and intercepts, we next investigated the use of modification indices to detect DIF.

Procedure. Following Yoon and Kim's (2014) recommendation, we used a sequential procedure to search modification indices for constraints to relax. For each simulated data set in conditions with DIF, we used the traditional or permuted $\Delta\chi^2$ to test the omnibus H_0. If we rejected the omnibus H_0, we calculated modification indices associated with freeing equality constraints on factor loadings. If the largest modification index was significant at the $\alpha = .05$ level, we specified a partial metric invariance model without that constraint and fit it to the data. If we still rejected the omnibus H_0 and the largest modification index from the less restricted model was also significant, we freed that constraint in a subsequent model. We repeated this process until the final model had either a nonsignificant omnibus test or no significant modification indices. We did this using each of three methods: $\Delta\chi^2$ with unadjusted *p* values for modification indices, $\Delta\chi^2$ with Bonferroni-adjusted *p* values for modification indices, and permuted $\Delta\chi^2$ with permutation-based Tukey-adjusted *p* values for modification indices.

Unadjusted *p* values were compared to $\alpha = .05$ at each iteration. Bonferroni-adjusted *p* values were equal to the unadjusted *p* values multiplied by the number of constraints tested at that iteration. For example, if there were six freely estimated factor loadings constrained to equality across two groups, then six modification indices would be calculated, so the unadjusted *p* value would be multiplied by six and then compared with $\alpha = .05$. The Tukey-adjusted *p* value for each constraint was calculated as the proportion of permutations that yielded a maximum modification index greater than the observed modification index, and Tukey-adjusted *p* values were also compared to $\alpha = .05$.

Analysis plan and hypotheses. For all models fit to each simulated data set, if DIF was falsely detected at least once for an indicator without DIF (e.g., any indicators loading on Factor 1), that replication was flagged as a false positive when calculating familywise Type I error rates. Likewise, if true DIF was detected for at least one indicator (e.g., Indicator 6 or 7 loading on Factor 2; see Table 1), then that replication was flagged as a true positive when calculating power. Multiple constraints could be freed for each simulated data set, so a single replication could possibly be flagged for both a false positive and a true DIF detection (or for neither).

Consistent with past research, we expected modification indices to have inflated familywise Type I error rates when using unadjusted *p* values (i.e., at least one Type I error would occur in more than 5% of replications when using $\alpha = .05$). Bonferroni-adjusted *p* values have been shown to adequately control Type I errors (French & Finch, 2008; Jak et al., 2010), and we expect Tukey-adjusted *p* values to control familywise Type I errors at least as well. Because the Bonferroni adjustment becomes more conservative as the number of tests increases, we expect the permutation-based Tukey adjustment to yield greater power, and for the discrepancy between these adjustments to increase with the number of tests.

Design. Because metric and scalar results were practically identically for the omnibus test, we restricted our investigation of modification indices to detect differences in factor loadings. The previous simulation studies showed that the omnibus $\Delta\chi^2$ and permutation tests have nominal Type I error rates when there is no DIF, so in this study we investigated familywise Type I error rates only when there was metric LOI (i.e., falsely detecting at least one valid equality constraint as invalid, when other equality constraints were truly invalid). Because the omnibus tests had such similar power to detect LOI when DIF = 0.3 and 0.4, we simulated only three levels of DIF (0.1, 0.2, and 0.3). We used the same range of sample sizes as all previous simulations, but we unlike other power analyses, we also manipulated model size to assess the effect of the number of tests on the discrepancy in power between the Bonferroni and Tukey adjustments. We used a 5 (*N*) × 3 (DIF) × 2 (four or eight indicators per factor) × 3 (*p* value adjustment method) design, holding the number of factors constant at two.

Results. The black lines in Figure 10 show only marginally greater power for unadjusted *p* values than for either Bonferroni

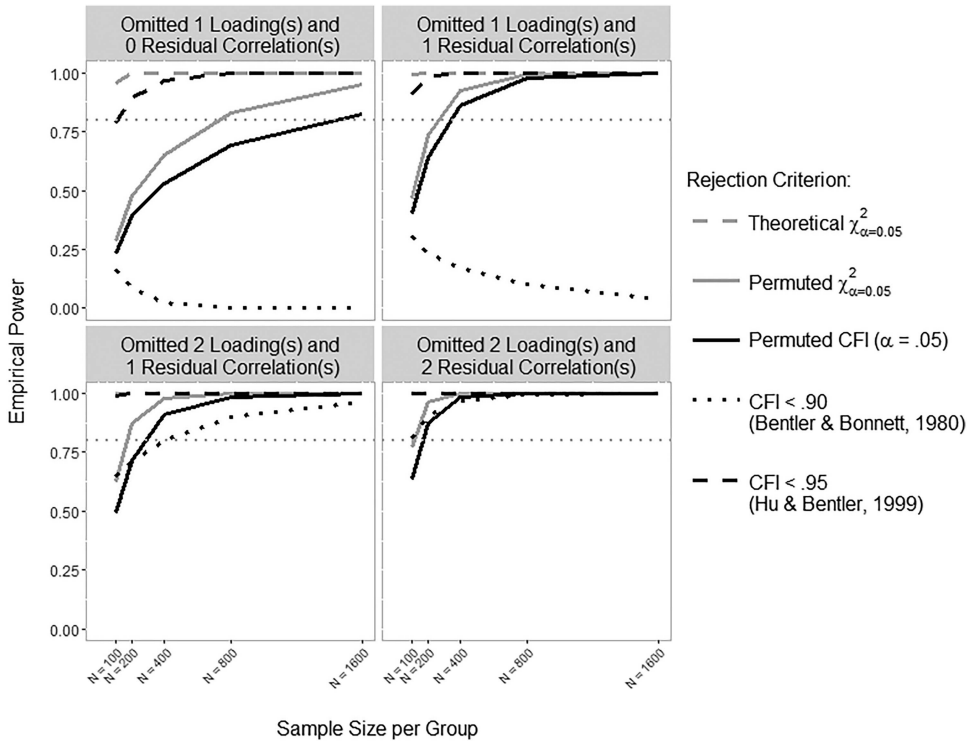

Figure 3. Power for χ^2 (gray lines) and CFI (black lines) using theoretical (or fixed) versus permutation-based critical values. The dotted gray line indicates 80% power.

or Tukey adjustments, but this came at the expense of inflated familywise Type I error rates (solid gray line). As expected, inflation with unadjusted *p* values became worse with more indicators to test, whereas Bonferroni and Tukey adjustments both yielded nominal familywise Type I error rates across conditions.

Contrary to our expectations, Bonferroni and Tukey adjustments yielded similar power to detect DIF for at least one indicator, irrespective of the number of indicators to test. It is difficult to distinguish between their dotted and dashed lines in Figure 10 because their rejection rates only differed in the third decimal place. Power was also nearly identical for both adjustments when calculating power separately for each differentially functioning indicator or power to detect all (rather than any) differentially functioning indicators (supplementary figures available in online supplemental materials). In general, adjusted *p* values had greater power when there were fewer indicators to test, which is not surprising given that both adjustments become more conservative as the number of tests increase.

Applied Example

We now provide an illustration of the permutation procedure for testing measurement equivalence using a real data set.

Method

The following example data are from Short and Hawley's (2015) investigation of college students' change in attitudes toward and knowledge of evolution across three different college courses with varying amounts of evolution education provided during the semester. Specifically, the Evolutionary Attitudes and Literacy Survey (EALS; Hawley, Short, McCune, Osman, & Little, 2011)—a multidimensional scale that consists of 16 subscales measuring an individual's knowledge of and attitudes toward evolution—was administered at the beginning and end of the semester to students enrolled in either a political science course that had no evolution education, a biology course that introduced the theory of evolution during the semester, or an evolutionary psychology course that consistently integrated the theory of evolution with psychological theories of human behavior throughout the entire semester. Although these data were longitudinal, we focused our illustration on just the first time point to evaluate group differences at the beginning of the semester. In addition, instead of examining DIF across all 104 indicators in the 16 subscales, we selected data from two subscales: young-earth creationism (six indicators) and intelligent design fallacies (six indicators; see Short & Hawley, 2012, for the list of indicators and additional details about participants). Thus, the examined model contained two factors with six indicators each and three groups: the political science course (n = 261), the biology course (n = 228), and the evolutionary psychology course (n = 63).

All models were estimated with maximum likelihood using lavaan (version 0.5–20; Rosseel, 2012). Models were identified with a standardized metric (i.e., latent variances fixed to one and latent means fixed to zero), and an appropriate null model (means

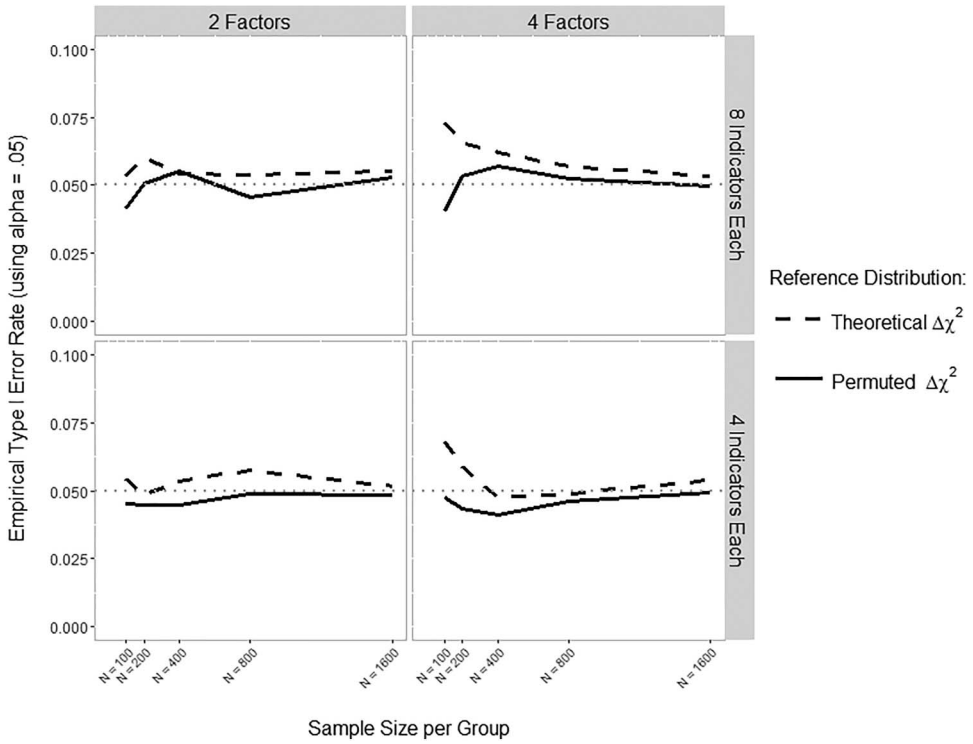

Figure 4. Type I error rates for $\Delta\chi^2$ and permutation test of metric invariance. The dotted gray line indicates the nominal error rate (5%). Note the *y*-axis ranges only from 0%–10%.

and variances constrained to equality across groups and all covariances fixed to zero; see Widaman & Thompson, 2003) was used for calculating CFI, which for simplicity is the only AFI we focus on in this illustration. For each test of invariance (i.e., configural, metric, and scalar), $I = 1{,}000$ permutations were used to calculate p values associated with $(\Delta)\chi^2$ and (Δ)CFI, using the R package semTools (semTools Contributors, 2016).

Results and Discussion

The two-factor configural model did not fit perfectly to the three groups, $\chi^2(159) = 443.93$, $p < .0001$, but the approximate fit was acceptable according to Bentler and Bonett (1980) CFI = .937, although not according to Hu and Bentler (1999). Seeking ways to improve the theoretical structure of the model is beyond the scope of this illustration, but despite the lack of perfect model fit, permutation tests using both $\chi^2(p = .54)$ and CFI ($p = .70$) supported configural ME/I across the three groups.

The metric invariant model, $\chi^2(179) = 388.682$, $p < .001$, CFI = .928, fit significantly worse than the configural model, $\Delta\chi^2(20) = 61.07$, $p < .001$, permuted $p = .001$, and approximate fit would only have been deemed similar to the configural model using Cheung and Rensvold's (2002) criterion (ΔCFI = −.009). However, permutation of ΔCFI did not support scalar invariance ($p = .001$). Using Tukey-adjusted p values based on the permutation distribution of the maximum modification index, modification indices suggested two indicators did not have equivalent factor loadings across all groups. We also report the expected parameter change (EPC) in each group; EPCs greater than 0.1 in absolute magnitude were deemed substantial. First, the young-earth creationism indicator "I read the bible literally" ($\lambda = 1.02$, $\Delta\chi^2 = 9.93$, $p = .03$) appeared to have a lower factor loading in the evolutionary psychology course (EPC = −0.38), a similar loading in the political science course (EPC = −0.02), and a higher loading in the biology course (EPC = 0.11). Second, the intelligent design fallacies indicator "Evolution is a theory in crisis" ($\lambda = 0.84$, $\Delta\chi^2 = 13.95$, $p = .003$) appeared to have a lower factor loading in the biology course (EPC = −0.167), a similar loading in the political science course (EPC = .05), and a higher loading in the evolutionary psychology course (EPC = 0.21). After freeing factor loadings for the first indicator, there was still indication of significant DIF in the second indicator, so those factor loadings were freed as well.

The partial metric invariant model, with factor loadings for these two indicators freely estimated, $\chi^2(175) = 475.53$, $p < .001$, CFI = .934, still fit significantly worse than the configural model, $\Delta\chi^2(16) = 31.60$, $p = .01$, permuted $p = .03$. Although the approximate fit would be deemed similar to configural using Cheung and Rensvold's (2002) or Chen's (2007) criteria (ΔCFI = −.003), the permuted ΔCFI showed significantly worse fit ($p = .03$). However, modification indices did not suggest any further evidence of DIF using Tukey-adjusted p values, so this partial metric invariance model was retained and examined for scalar invariance.

The scalar invariant model constrained all intercepts to equality, except for the intercepts of the two indicators whose load-

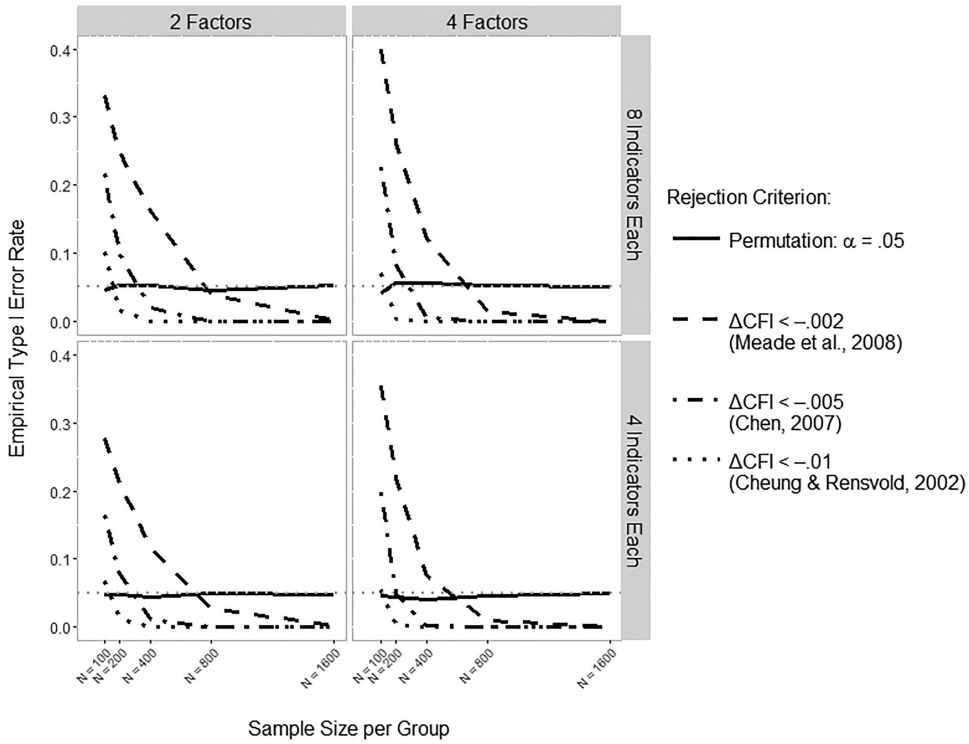

Figure 5. Type I error rates for ΔCFI test of metric invariance, using fixed and permutation-based critical values. The dotted gray line indicates the nominal error rate (5%). Note the *y*-axis ranges only from 0%–40%.

ings were unconstrained, $\chi^2(191) = 514.88$, CFI = .929. The scalar model fit significantly worse than the partial metric model, $\Delta\chi^2(16) = 39.36$, $p < .001$, permuted $p = .002$. Approximate fit was only similar to the partial metric model according to Cheung and Rensvold's (2002) criterion (ΔCFI = −.005), but the permutation of ΔCFI did not support scalar invariance ($p = .003$). Modification indices using Tukey-adjusted p values suggested the intercepts of two indicators of intelligent design fallacies might differ across groups. First, the indicator "there is scientific evidence that humans were created by a supreme being or intelligent designer" ($\tau = 3.40$, $\Delta\chi^2 = 16.50$, $p = .003$) appeared to have a lower indicator intercept for the evolutionary psychology course (EPC = −0.34), a similar intercept for the political science course (EPC = −0.09), and a higher intercept for the biology course (EPC = 0.34). Second, the indicator "there are no transitional fossils (remains of life forms that illustrate an evolutionary transition)" ($\tau = 2.93$, $\Delta\chi^2 = 10.24$, $p = .03$) appeared to have a lower indicator intercept for the biology course (EPC = −0.18), but higher indicator intercepts for the political science (EPC = 0.11) and evolutionary psychology (EPC = 0.13) courses.

Because both indicators were indicators of the same construct, freeing one intercept across groups would likely change the estimates of other intercepts. Thus, we fit a partial scalar invariance model by freeing the intercepts across groups only for the first indicator described, which had the larger modification index and EPCs, leaving the equality constraint intact for the second indicator, $\chi^2(189) = 495.57$, $p < .001$, CFI = .932. This partial scalar model did not fit significantly worse than the partial metric model, $\Delta\chi^2(14) = 20.04$, $p = .15$, permuted $p = .13$, and the approximate fit was similar, ΔCFI = −.001 (permuted $p = .14$). This model was therefore retained, and no further modifications were examined.

General Discussion

Summary of Empirical Results

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 χ^2 test 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 (Δ)AFIs over (Δ)χ^2 (Putnick & Bornstein, 2016) because of the latter's sensitivity to differences that are negligible in practice, which could be thought of as

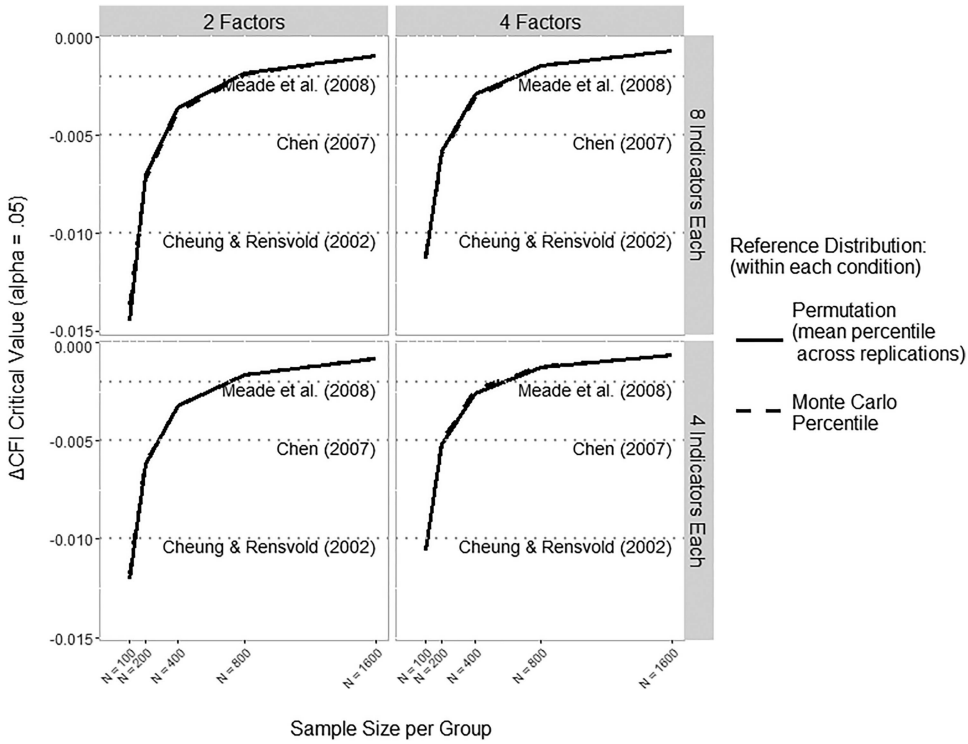

Figure 6. Critical values for ΔCFI test of metric invariance, based on Monte Carlo 95th percentile in each condition and on the average permutation 95th percentile in each condition. The dotted gray lines indicate previously suggested fixed critical values.

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 $\Delta\chi^2$.

The third limitation we proposed to address with permutation is that after rejecting full ME/I, testing multiple indicators for DIF leads to inflated familywise Type I error rates. Although a Bonferroni correction provides adequate Type I error control (French & Finch, 2008; Jak et al., 2010), we conceived of a method similar to Tukey's HSD, which we anticipated would provide similar control but greater power than a Bonferroni correction. However, our simulations showed that the permutation-based Tukey method provided nearly identical power as well as familywise Type I error rates, regardless of the number of tests conducted.

Recommendations

We recommend that applied researchers interested in testing configural invariance use the permutation method. 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. Two situations illustrate why relying on absolute measures of model fit can lead to an incorrect conclusion. It is possible that the H_0 of group equivalence is true even for a poorly fitting model, in which case the model should be rejected because it is a poor approximation of at least one group's data-generating process, not necessarily because of group differences. Another possibility is that the group models substantially differ in functional form, but that a simpler model fits each population well enough to result in population-level AFIs that indicate acceptable approximate fit. In such a case, the H_0 would actually be false, but there would be little chance of detecting the lack of configural invariance (especially for AFIs).

We do not recommend that researchers assessing more restrictive levels of ME/I use fixed cutoffs for ΔAFIs because they lead to inflated Type I error rates at smaller sample sizes. Even when sample sizes are large enough to have low Type I errors, ΔAFIs have no power advantage over permutation or $\Delta\chi^2$. We recommend permutation or $\Delta\chi^2$ because both have nominal Type I error rates and similar power across conditions. When locating differentially functioning indicators after rejecting full ME/I, we recommend using modification indices only with a Bonferroni adjustment or, if already permuting, using the permutation distribution of the maximum modification index to calculate *p* values. These methods have similar power to an unadjusted significant criterion, but they keep the familywise Type I error rate nominal.

Software

The permutation method is implemented in the R package semTools, using the function "permuteMeasEq" (semTools

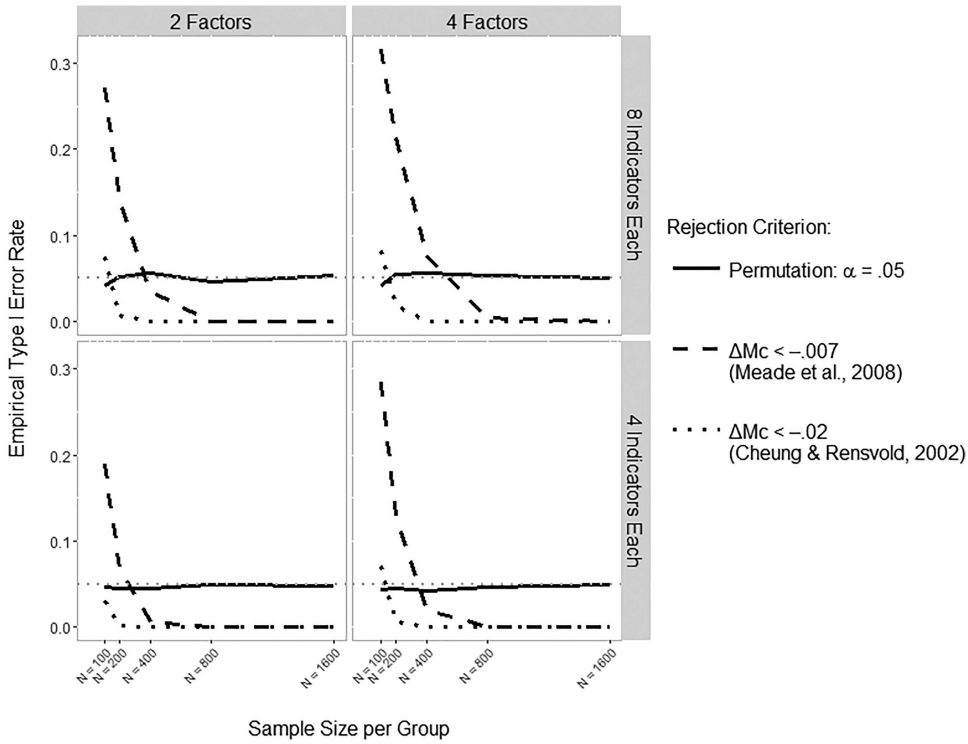

Figure 7. Type I error rates for ΔMc test of metric invariance, using fixed and permutation-based critical values. The dotted gray line indicates the nominal error rate (5%). Note the *y*-axis ranges only from 0%–40%.

Contributors, 2016). Users can fit their multiple-Group CFA models using the lavaan (Rosseel, 2012) software. Users can then submit their lavaan models to the "permuteMeasEq" function and specify additional information, such as the number of permutations, which fit measure to permute, and which parameters are being tested for equality. Examples of using this software can be found by accessing the help page for "permuteMeasEq" within R. The supplementary online materials also

Table 4
Difference in Fit Measures Between Metric and Configural Models Fit to Population Data When H_0 Is False

		DIF			
Fit measure	*N* per group	.1	.2	.3	.4
$\Delta\chi^2_{df=6}$ (power)	100	1.23 (11%)	4.72 (32%)	10.08 (65%)	16.93 (89%)
	200	2.46 (17%)	9.43 (61%)	20.16 (94%)	33.85 (100%)
	400	4.93 (33%)	18.87 (92%)	40.32 (100%)	67.71 (100%)
	800	9.86 (64%)	37.73 (100%)	80.64 (100%)	135.41 (100%)
	1,600	19.72 (94%)	75.46 (100%)	161.27 (100%)	270.82 (100%)
ΔCFI	100	0	0	0	0
	200	0	0	0	0
	400	0	0	−.007	−.020
	800	−.0002	−.007	−.018	−.031
	1,600	−.002	−.008	−.019	−.032
ΔMc	100	.013	.003	−.011	−.029
	200	.005	−.004	−.018	−.035
	400	.001	−.008	−.021	−.038
	800	−.001	−.010	−.023	−.040
	1,600	−.002	−.011	−.024	−.040
ΔRMSEA	100	0	0	0	0
	200	0	0	0	0
	400	0	0	.028	.048
	800	.005	.029	.045	.060
	1,600	.003	.015	.029	.042
ΔSRMR	≤1,600	.005	.016	.028	.041

Note. Power for $\Delta\chi^2$ calculated using method described in Satorra and Saris (1985).

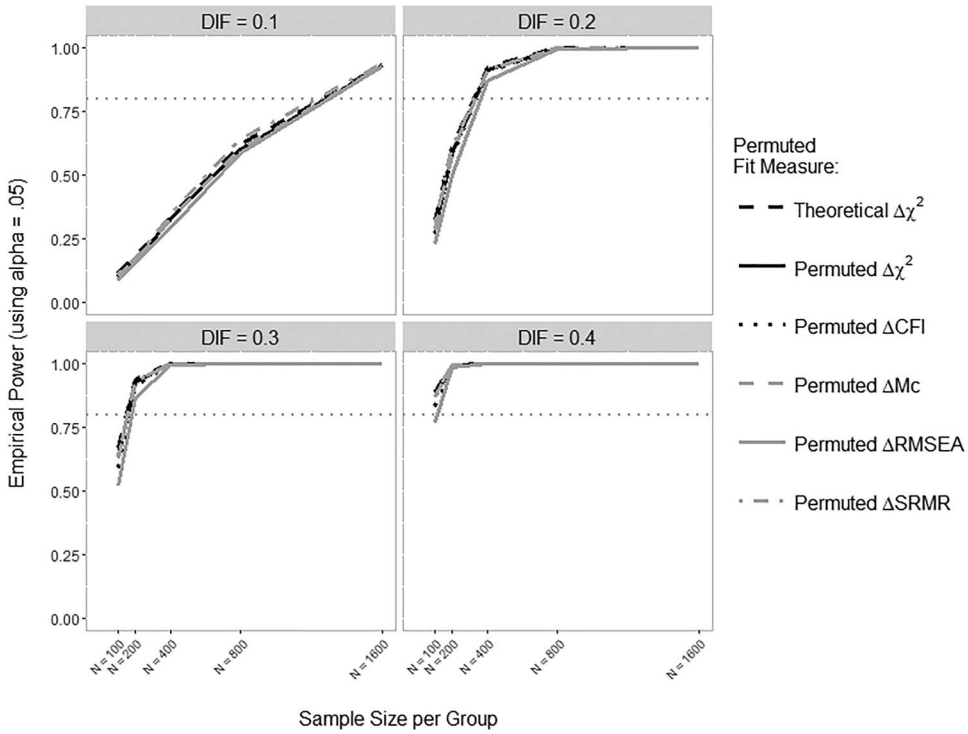

Figure 8. Power for test of metric invariance using traditional $\Delta\chi^2$, permuted $\Delta\chi^2$, and permuted ΔAFIs. The dotted gray line indicates power = 80%.

include the R syntax we used for the Applied Example section.[17]

Users who prefer to use other dedicated SEM software could still benefit from R (or other software such as Fortran, C++, or Python), such as using R's "sample" function to permute group assignment, saving a large set of permuted data sets to be analyzed outside of R. Syntax files for the SEM software would still have to be written to analyze each of the permuted data sets, at which point R could also be used to aggregate results from the resulting output files. The supplementary online materials include R syntax demonstrating how to implement the permutation method using Mplus (Muthén & Muthén, 2012), which is facilitated greatly by the R package MplusAutomation (Hallquist & Wiley, 2016), as well as by the MONTECARLO feature in Mplus.

Limitations and Future Directions

Although proposed cutoffs for (Δ)AFIs were only meant to be rules of thumb (Cheung & Rensvold, 2002; Hu & Bentler, 1999), applied researchers may be inclined to treat cutoffs as critical values (Chen, 2007; Putnick & Bornstein, 2016), effectively treating a descriptive index of (change in) model fit as though it were a test statistic. Our goal was not to argue how AFIs should be treated, but rather to provide researchers who want to treat (Δ)AFIs as test statistics with a method that yields nominal Type I error rates. We recognize that researchers who prefer (Δ)AFIs use them in lieu of (Δ)χ^2 precisely because of the test statistic's sensitivity to smaller (perhaps negligible) DIF in larger samples. Permuting (Δ)AFIs does not solve this problem because they are just as sensitive as (Δ)χ^2, so other solutions must be considered.

For example, Oberski (2014) advocated addressing ME/I by focusing not on the equivalence of measurement parameters but on the research questions that assume ME/I. In our applied example using Short and Hawley's (2015) data, the ultimate goal was to compare latent means across groups, which would require at least partial scalar ME/I. When we found evidence against full metric and scalar ME/I, we considered EPCs in tandem with modification indices (Whittaker, 2012) to estimate how much a differentially functioning indicator's parameter estimates would change if the same estimates were freed. We could also have considered effect sizes as proposed by Millsap and Olivera-Aguilar (2012).[18] But because the latent-mean comparisons were of primary interest, we could instead have used the *EPC-interest* (Oberski, 2014) to estimate the expected change in latent mean estimates if an indicator's intercepts were freed. Substantial expected latent-mean changes could be used to flag the particular indicator whose constraints should be freed. This method might benefit from permutation, which could be used to estimate a distribution of the maximum EPC-interest under the H_0, thus controlling familywise Type I

[17] We do not provide the original data, but we do provide a simulated data set that resembles the real data we analyzed. Therefore, running the syntax from the online supplemental materials will not yield the exact same results as reported.

[18] Millsap and Olivera-Aguilar (2012) discussed effect sizes in terms of continuous indicators, for example, expressing the proportion of group mean differences in observed indicators that is due to differences in intercepts rather than differences in common-factor means. With respect to categorical indicators, Meade (2010) provided a taxonomy of various effect size measures for DIF in the IRT context.

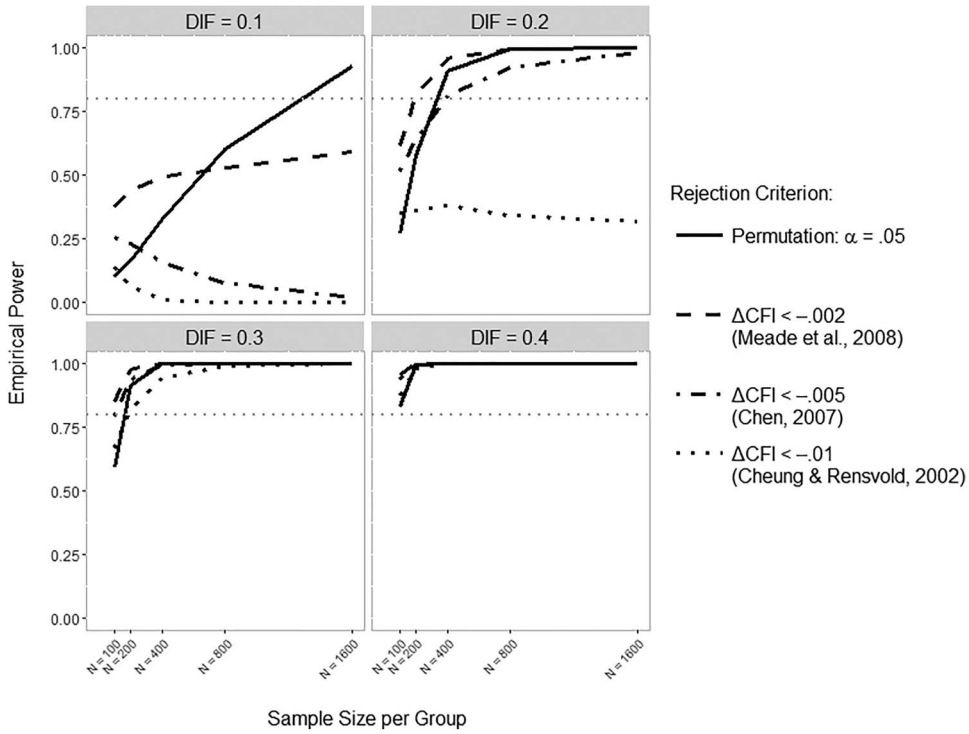

Figure 9. Power for ΔCFI test of metric invariance using permutation and fixed cutoffs. The dotted gray line indicates the power = 80%.

errors. The "permuteMeasEq" function allows users to estimate the permutation distribution of any statistic of interest, which should facilitate future research on this topic.

Consistent with the Meade et al. (2008) study on which our simulations were largely based, we used a reference indicator without DIF in our analysis models. Thus, our omnibus H_0 results would be identical if we had identified the models by fixing all factor variances to one in the configural model, and freeing the factor variances in the second group when we constrained loadings to equality. In practice, we recommend researchers fix factor variances instead of factor loadings, so that they need not assume any particular variable functions equivalently across groups without testing that assumption. However, when multiple indicators have DIF, freeing one indicator at a time to establish partial ME/I can inflate Type I error rates because misspecification can bias other measurement parameters (French & Finch, 2008; Woods, 2009[19]). Due to Meade et al.'s (2008) study design, we found no inflation because each indicator with DIF had a corresponding indicator within the same factor that had the same magnitude of DIF, but in the opposite direction. Because permutation eliminates group differences in all indicators, it might advantageously free us from the assumption that all other indicators are invariant when testing any one indicator; however, our use of Meade et al.'s (2008) study design prevented us from testing this in the current study. We encourage further investigation.

Given the additional computation time needed to permute the data, there is no apparent justification for permuting instead of using $\Delta\chi^2$ when testing any level of ME/I other than configural. However, we simulated rather ideal circumstances: complete data, balanced group sizes, group differences only in measurement parameters, multivariate normality. There are certain conditions in which χ^2 yields inflated Type I error rates, such as using full-information maximum likelihood estimation on small multivariate normal samples with missing data (Savalei & Bentler, 2009) or using weighted least squares estimation with ordinal indicators (Sass, Schmitt, & Marsh, 2014), particularly when thresholds are asymmetric (Bandalos, 2014). In these situations (similar to the case of configural ME/I when the model does not fit perfectly), any noncentrality in the χ^2 distribution should be preserved in the permuted data, so the permutation distribution under the H_0 of group equivalence should maintain nominal Type I error rates better than a test statistic whose sampling distribution is not adequately characterized by the central χ^2 distribution. Further research will be needed to establish whether this hypothesis is supported.

Permutation methods might also be expected to perform poorly under certain conditions. Although the permutation test may not assume a specific distribution of the data, it does assume exchangeability of observations (Hayes, 1996). When data are not randomized across groups, exchangeability requires that each group's distribution is the same shape. If common or unique factor variances are heteroscedastic across groups, permuting the groups would create a mixture of distributions with different variances, probably affecting covariance-structure estimates such as factor

[19] French and Finch (2008) and Woods (2009) also discuss and evaluate empirical methods to select sets of invariant indicators.

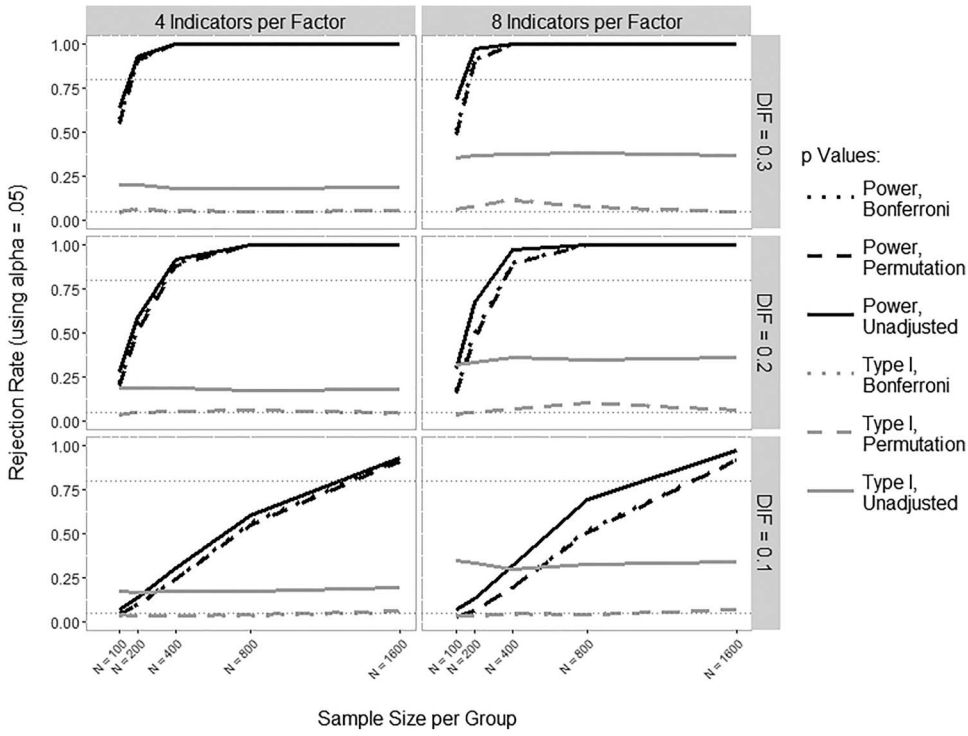

Figure 10. Familywise Type I error rates (gray lines) and power (black lines) of modification indices to detect metric LOI, using unadjusted (solid), Bonferroni-adjusted (dotted), and permutation-based Tukey-adjusted (dashed) p values. Thin dotted reference lines are provided at 5% and 80% probability of rejecting H_0.

loadings. The permutation method proposed here is based on model fit measures rather than parameter estimates, so it may be robust to the exchangeability assumption. Future research should investigate this robustness issue.

Conclusion

Permutation randomization is a potentially valuable method for testing ME/I across groups. It provides a valid test of configural invariance, and provides well controlled Type I error rates across a variety of conditions, regardless of a researcher's preferred fit measure. 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, not only using multigroup CFA but also IRT and multiple-indicator multiple-cause (MIMIC) models.

References

Asparouhov, T., & Muthén, B. O. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling, 21,* 495–508. http://dx.doi.org/10.1080/10705511.2014.919210

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

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

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

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

Brannick, M. T. (1995). Critical comments on applying covariance structure modeling. *Journal of Organizational Behavior, 16,* 201–213. http://dx.doi.org/10.1002/job.4030160303

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York, NY: Guilford Press.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21,* 230–258. http://dx.doi.org/10.1177/0049124192021002005

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105,* 456–466. http://dx.doi.org/10.1037/0033-2909.105.3.456

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

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology, 95,* 1005–1018. http://dx.doi.org/10.1037/a0013193

Cheung, G. W., & Lau, R. S. (2012). A direct comparison approach for testing measurement invariance. *Organizational Research Methods, 15,* 167–198. http://dx.doi.org/10.1177/1094428111421987

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9,* 233–255. http://dx.doi.org/10.1207/S15328007SEM0902_5

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the "problem" of sample size: A clarification. *Psychological Bulletin, 109,* 512–519. http://dx.doi.org/10.1037/0033-2909.109.3.512

French, B. F., & Finch, W. H. (2008). Multigroup confirmatory factor analysis: Locating the invariant referent sets. *Structural Equation Modeling, 15,* 96–113. http://dx.doi.org/10.1080/10705510701758349

Gonzalez-Roma, V., Tomas, I., Ferreres, D., & Hernandez, A. (2005). Do items that measure self-perceived physical appearance function differentially across gender groups? An application of the MACS model. *Structural Equation Modeling, 12,* 148–162. http://dx.doi.org/10.1207/s15328007sem1201_8

Hallquist, M., & Wiley, J. (2016). MplusAutomation: Automating Mplus model estimation and interpretation (version 0.6–4) [Computer software]. Retrieved from: https://www.R-project.org/

Hawley, P. H., Short, S. D., McCune, L. A., Osman, M. R., & Little, T. D. (2011). What's the matter with Kansas? The development and confirmation of the Evolutionary Attitudes and Literacy Survey (EALS). *Evolution: Education and Outreach, 4,* 117–132.

Hayes, A. F. (1996). Permutation test is not distribution-free: Testing H_0: $\rho = 0$. *Psychological Methods, 1,* 184–198. http://dx.doi.org/10.1037/1082-989X.1.2.184

Higgins, J. J. (2004). *Introduction to modern nonparametric statistics*. Pacific Grove, CA: Duxbury Press.

Hildreth, L. A., Genschel, U., Lorenz, F. O., & Lesser, V. M. (2013). A permutation test for correlated errors in adjacent questionnaire items. *Structural Equation Modeling, 20,* 226–240. http://dx.doi.org/10.1080/10705511.2013.769390

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55. http://dx.doi.org/10.1080/10705519909540118

Jak, S., Oort, F. J., & Dolan, C. V. (2010). Measurement bias and multidimensionality: An illustration of bias detection in multidimensional measurement models. *Advances in Statistical Analysis, 94,* 129–137. http://dx.doi.org/10.1007/s10182-010-0128-z

Jorgensen, T. D., Kite, B., Chen, P.-Y., & Short, S. D. (2017). Finally! A valid test of configural invariance using permutation in multigroup CFA. In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W.-C. Wang (Eds.), *Quantitative psychology: The 81st annual meeting of the psychometric society, Asheville, NC, 2016*. New York, NY: Springer. http://dx.doi.org/10.1007/978-3-319-56294-0_9

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research, 44,* 486–507. http://dx.doi.org/10.1177/0049124114543236

Kim, E. S., & Willson, V. L. (2014). Testing measurement invariance across groups in longitudinal data: Multigroup second-order latent growth model. *Structural Equation Modeling, 21,* 566–576. http://dx.doi.org/10.1080/10705511.2014.919821

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.

Linting, M., van Os, B. J., & Meulman, J. J. (2011). Statistical significance of the contribution of variables to the PCA solution: An alternative permutation strategy. *Psychometrika, 76,* 440–460. http://dx.doi.org/10.1007/s11336-011-9216-6

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling, 13,* 59–72. http://dx.doi.org/10.1207/s15328007sem1301_3

MacCallum, R. C. (2003). 2001 presidential address: Working with imperfect models. *Multivariate Behavioral Research, 38,* 113–139. http://dx.doi.org/10.1207/S15327906MBR3801_5

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

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111,* 490–504. http://dx.doi.org/10.1037/0033-2909.111.3.490

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). 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. *Structural Equation Modeling, 11,* 320–341. http://dx.doi.org/10.1207/s15328007sem1103_2

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification, 6,* 97–103. http://dx.doi.org/10.1007/BF01908590

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology, 95,* 728–743. http://dx.doi.org/10.1037/a0018966

Meade, A. W., & Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling, 14,* 611–635. http://dx.doi.org/10.1080/10705510701575461

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. http://dx.doi.org/10.1037/0021-9010.93.3.568

Meade, A. W., & Lautenschlager, G. J. (2004). A Monte-Carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling, 11,* 60–72. http://dx.doi.org/10.1207/S15328007SEM1101_5

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

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Millsap, R. E., & Kwok, O.-M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods, 9,* 93–115. http://dx.doi.org/10.1037/1082-989X.9.1.93

Millsap, R. E., & Olivera-Aguilar, M. (2012). Investigating measurement invariance using confirmatory factor analysis. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 380–392). New York, NY: Guilford Press.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

Nevitt, J., & Hancock, G. R. (2004). Evaluating small sample approaches for model test statistics in structural equation modeling. *Multivariate Behavioral Research, 39,* 439–478. http://dx.doi.org/10.1207/S15327906MBR3903_3

Oberski, D. L. (2014). Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Political Analysis, 22,* 45–60. http://dx.doi.org/10.1093/pan/mpt014

Pornprasertmanit, S., Wu, W., & Little, T. D. (2013). Using a Monte Carlo approach for nested model comparisons in structural equation modeling. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), *New developments in quantitative psychology* (pp. 187–197). New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4614-9348-8_12

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

R Core Team. (2016). R: A language and environment for statistical computing (version 3.3.0) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rodgers, J. L. (1999). The bootstrap, the jackknife, and the randomization test: A sampling taxonomy. *Multivariate Behavioral Research, 34,* 441–456. http://dx.doi.org/10.1207/S15327906MBR3404_2

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

Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling, 21,* 167–180. http://dx.doi.org/10.1080/10705511.2014.882658

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

Savalei, V., & Bentler, P. M. (2009). A two-stage approach to missing data: Theory and application to auxiliary variables. *Structural Equation Modeling, 16,* 477–497. http://dx.doi.org/10.1080/10705510903008238

semTools Contributors. (2016). semTools: Useful tools for structural equation modeling (version 0.4–12) [Computer software]. Retrieved from https://www.R-project.org/

Short, S. D., & Hawley, P. H. (2012). Evolutionary Attitudes and Literacy Survey (EALS): Development and validation of a short form. Evolution. *Education and Outreach, 5,* 419–428. http://dx.doi.org/10.1007/s12052-012-0429-7

Short, S. D., & Hawley, P. H. (2015). The effects of evolution education: Examining attitudes toward and knowledge of evolution in college courses. *Evolutionary Psychology, 13,* 67–88. http://dx.doi.org/10.1177/147470491501300105

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *The Journal of Consumer Research, 25,* 78–107. http://dx.doi.org/10.1086/209528

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

Tueller, S. J., Drotar, S., & Lubke, G. H. (2011). Addressing the problem of switched class labels in latent variable mixture model simulation studies. *Structural Equation Modeling, 18,* 110–131. http://dx.doi.org/10.1080/10705511.2011.534695

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

Whittaker, T. A. (2012). Using the modification index and standardized expected parameter change for model modification. *Journal of Experimental Education, 80,* 26–44. http://dx.doi.org/10.1080/00220973.2010.531299

Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods, 8,* 16–37. http://dx.doi.org/10.1037/1082-989X.8.1.16

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement, 33,* 42–57. http://dx.doi.org/10.1177/0146621607314044

Yoon, M., & Kim, E. S. (2014). A comparison of sequential and nonsequential specification searches in testing factorial invariance. *Behavior Research Methods, 46,* 1199–1206. http://dx.doi.org/10.3758/s13428-013-0430-2

Received August 12, 2016
Revision received May 2, 2017
Accepted May 13, 2017 ■