Random Permutation Testing Applied to Measurement Invariance Testing with Ordered-Categorical Indicators

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

DOI
10.1080/10705511.2017.1421467

Publication date
2018

Document Version
Final published version

Published in
Structural Equation Modeling

Routledge
Taylor & Francis Group

Structural Equation Modeling: A Multidisciplinary Journal

ISSN: 1070-5511 (Print) 1532-8007 (Online) Journal homepage: https://www.tandfonline.com/loi/hsem20

Random Permutation Testing Applied to Measurement Invariance Testing with Ordered-Categorical Indicators

Benjamin A. Kite, Terrence D. Jorgensen & Po-Yi Chen

To cite this article: Benjamin A. Kite, Terrence D. Jorgensen & Po-Yi Chen (2018) Random Permutation Testing Applied to Measurement Invariance Testing with Ordered-Categorical Indicators, Structural Equation Modeling: A Multidisciplinary Journal, 25:4, 573-587, DOI: 10.1080/10705511.2017.1421467

To link to this article: https://doi.org/10.1080/10705511.2017.1421467

Published online: 24 Jan 2018.

Structural Equation Modeling: A Multidisciplinary Journal, 25: 573–587, 2018

ISSN: 1070-5511 print / 1532-8007 online
DOI: https://doi.org/10.1080/10705511.2017.1421467

Random Permutation Testing Applied to Measurement Invariance Testing with Ordered-Categorical Indicators

Benjamin A. Kite,[1] Terrence D. Jorgensen,[2] and Po-Yi Chen[1]
[1]*University of Kansas*
[2]*University of Amsterdam*

We describe and evaluate a random permutation test of measurement invariance with ordered-categorical data. To calculate a *p*-value for the observed $(\Delta)\chi^2$, an empirical reference distribution is built by repeatedly shuffling the grouping variable, then saving the χ^2 from a configural model, or the $\Delta\chi^2$ between configural and scalar-invariance models, fitted to each permuted dataset. The current gold standard in this context is a robust mean- and variance-adjusted $\Delta\chi^2$ test proposed by Satorra (2000), which yields inflated Type I errors, particularly when thresholds are asymmetric, unless samples sizes are quite large (Bandalos, 2014; Sass et al., 2014). In a Monte Carlo simulation, we compare permutation to three implementations of Satorra's robust χ^2 across a variety of conditions evaluating configural and scalar invariance. Results suggest permutation can better control Type I error rates while providing comparable power under conditions that the standard robust test yields inflated errors.

Keywords: measurment invariance, permutation testing, categorical indicators

The goal of this article is to propose a permutation test of the omnibus null hypothesis of measurement equivalence/invariance in confirmatory factor analysis (CFA) models fit to discrete indicators, sometimes referred to as *item factor analysis* (Wirth & Edwards, 2007). Before we describe the permutation testing procedure—first introduced by Jorgensen, Kite, Chen, and Short (2017a) in the context of multivariate normal indicators—we briefly review the common-factor model for discrete data, discuss relevant estimation and inference problems and their solutions, and review evidence from background literature indicating how well these solutions perform in practice. Based on previous results, we hypothesize under what conditions we expect permutation to control Type I error rates better than the current gold standard. Motivated by these hypotheses, we describe a simulation study that compares permutation to the current gold standard. Following an illustrative analysis of real data, we conclude with advice for applied researchers and future methodological research.

Correspondence should be addressed to Benjamin A. Kite, University of Kansas, Lawrence, KS 66045. E-mail: bakite@ku.edu

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hsem.

ESTIMATING AND TESTING CFA MODELS WITH ORDERED-CATEGORICAL DATA

Maximum likelihood estimation algorithms available in most structural equation modeling (SEM) software programs assume that data are multivariate normally distributed, but social and behavioral scientists frequently utilize measurement scales consisting of a few ordinal categories, for example, 0 (*never*), 1 (*seldom*), 2 (*often*), or 3 (*always*). Standard errors and test statistics can be adjusted for non-normality to keep Type I error rates closer to the nominal level, but point estimates remain attenuated unless scale items include at least five or seven categories, depending on the distribution of thresholds (Rhemtulla, Brosseau-Liard, & Savalei, 2012).

Advances in item factor analysis (Jöreskog & Moustaki, 2001; Muthén, 1984; Wirth & Edwards, 2007) have allowed factor analysis models to be fit to ordinal indicators by including a threshold model, which is statistically equivalent to some item-response theory models (Kamata & Bauer, 2008; Wirth & Edwards, 2007). The threshold model is predicated on the assumption that a normally distributed latent item-response underlies each observed discrete indicator. Thresholds are values on the scale of that latent item-

response (e.g., assuming a standard normal distribution, thresholds are z scores) that delimits adjacent categories of the observed response scale. Thus, the threshold model implies that when a respondent's latent item-response exceeds a threshold, the respondent's observed response is in the higher rather than the lower category. Based on this simplifying assumption, thresholds are estimated using the observed proportions in each response category to indicate which corresponding z scores would yield those cumulative probabilities in a standard normal distribution, and polychoric correlations[1] can be estimated among discrete items based on those thresholds. The common-factor model can then be fitted to those polychoric correlations.

This two-step estimation process requires accounting for the uncertainty about estimated polychoric correlations when estimating structural parameters that explain those correlations. A weighted least-squares (WLS) estimation algorithm has been adapted (Muthén, 1984) specifically for this scenario, which incorporates the asymptotic covariance matrix of estimated polychoric correlations in Step 1 as the weight matrix W in Step 2. Estimates for the CFA parameters are obtained by minimizing the discrepancy function F_{WLS}.

$$F_{WLS} = (s - \sigma)^T W^{-1} (s - \sigma)$$

where s is the vector of estimated polychoric correlations from Step 1 and σ is the vector of polychoric correlations implied by the hypothesized model fit in Step 2. WLS estimation becomes computationally intensive in practice when the number of variables (p) becomes quite large. For example, fitting a CFA model to only $p = 10$ indicators already requires estimating $p^* = p(p-1)/2 = 45$ polychoric correlations, in which case W would be a 45×45 symmetric matrix. Inverting large matrices can be prohibitively intensive with many variables.

Alternatively, estimates can be obtained by using only the diagonal of W, which is much easier to invert. This is commonly referred to as diagonally weighted least squares (DWLS) estimation, which yields asymptotically unbiased point estimates but biased *SE*s and test statistics. Our focus in this article is on the χ^2 model-fit statistic, so we will not discuss robust *SE*s. Among the most popular robust corrections to the χ^2 model-fit statistic when using DWLS is a mean- and variance-adjusted χ^2 statistic[2] (χ^2_{MV}) proposed by Satorra (2000), which is widely implemented in SEM software, although implementations can differ across packages (WirthEdward; see below for details). Whereas WLS requires very large sample sizes (e.g., $N > 2000$) for the χ^2 model-fit statistic to yield nominal Type I error rates, with more moderate sizes (e.g., $200 < N < 1000$) the robust χ^2_{MV} under DWLS estimation yields Type I error rates closer to the nominal level than χ^2 under full WLS (Flora & Curran, 2004). However, Bandalos (2014) showed that even when samples are moderately large, χ^2_{MV} under DWLS will return inflated Type I errors when thresholds are not approximately symmetric. The less symmetric thresholds are the fewer cases will be observed in extreme categories that represent the tails of the distribution of the latent item-response. Thus, unless the sample size is large enough so that the probabilities in each response category can be reliably estimated, results based on imprecisely estimated thresholds will be biased.

When comparing nested models estimated with DWLS, the $\Delta\chi^2$ statistic[3] also requires a robust correction. The mean-and-variance adjustment proposed by Satorra (2000), as implemented in M*plus* (Muthén & Muthén, 2015) using the DIFFTEST procedure, also appears to yield nearly nominal Type I error rates, at least when samples are adequately large (N = 1100 or 2000; Asparouhov & Muthén, 2006). However, Sass, Schmitt, and Marsh (2014) found the same problems with the empirical performance of $\Delta\chi^2_{MV}$ that Bandalos (2014) found with χ^2_{MV}. Namely, asymmetric thresholds inflate Type I error rates, and the problem is exacerbated by small sample size.

The open-source statistical environment R (R Core Team, 2017) affords an opportunity for users to directly investigate different implementations of the $(\Delta)\chi^2_{MV}$ solution in the R package lavaan (Rosseel, 2012). When comparing nested models fitted with DWLS estimation in lavaan, $\Delta\chi^2_{MV}$ can be requested using the lavTestLRT function. There are two alternative calculations available via the method argument, which specifies how the Jacobian of the constraint function is to be computed (Satorra, 2000, p. 239) before it is used in their Equations 22 (p. 241) and 23 (p. 242). The first option (method = "exact") requests an exact solution, which requires finding a function that can be imposed on the parameter vector of the less restricted model that reflects the (additional) constraints imposed in the more restricted model. If all constraints were satisfied, this function would return a matrix of zeroes. The exact solution therefore requires that the two models have nested parameter vectors, in that the parameters of the more restricted model are a subset of the parameters of the less restricted model. The second option (method = "delta") approximates the Jacobian

[1] Throughout the manuscript, we will restrain our discussion to the case of polychoric correlations for models fit only to ordered-categorical items, but this WLS estimator can also be applied to a mixture of discrete and continuous indicators. When continuous indicators are included, their observed (co)variances are included in the estimated polychoric correlation matrix, and polyserial correlations are estimated between the discrete and continuous indicators.

[2] Mean- and variance-adjusted χ^2 statistics can also be calculated for other estimators, such as maximum likelihood.

[3] Note that it is not appropriate to calculate the difference between two χ^2_{MV} statistics because they will not be approximately χ^2 distributed. Instead, the difference between unadjusted χ^2 statistics must be calculated, then adjusted.

as the orthogonal complement to the following function of the derivatives of model parameters from the restricted (D_0) and unrestricted (D_1) models:

$$(D_1^T D_1)^{-1} D_1^T D_0$$

This approximation does not necessarily result in a constraint matrix of zeroes, but it is more flexible because it merely requires nested covariance structures rather than strict parameter nesting (Bentler & Satorra, 2010), such that the set of all predictions that could possibly be made by the parent model include all possible predictions made by the nested model. The exact option was the default in lavaan versions 0.5 (the first to include DWLS estimation for categorical indicators), but since version 0.6–1.1109 the new default method is the approximate delta method. To the best of our knowledge, these methods have not been compared empirically.

Other open-source SEM software packages have also implemented $\Delta\chi^2_{MV}$, such as the R package OpenMx (Neale et al., 2016). Proprietary SEM software, however, is not always transparent about how solutions like this are implemented. For example, the DIFFTEST[4] command in M*plus* (Muthén & Muthén, 1998–2015) implements the Satorra (2000) correction, but it is unclear how M*plus* calculates the Jacobian of the constraint function (Asparouhov & Muthén, 2006, eq. 6). A comparison with the open-source solutions in lavaan could provide insight about which approach M*plus* appears to use (if either), which might be of practical interest if the methods yield substantially different frequency properties (e.g., Type I error rates, power) under different realistic conditions (e.g., when sample size is not large enough to rely on asymptotic assumptions).

TESTING MEASUREMENT INVARIANCE WITH ORDERED-CATEGORICAL DATA

Measurement invariance in a CFA framework is traditionally tested beginning with a baseline multi group model that assumes equivalent model configurations across groups (i.e., same specification of fixed and free parameters) but places no restrictions on the estimated values of those parameters across groups (Vandenberg & Lance, 2000). The fit of this configural baseline model is evaluated using a χ^2 test of exact fit, where a rejection of the null hypothesis[5] implies that the configural model does not exactly represent the true data-generating process (i.e., population). In practice, researchers frequently rely on alternative fit indices (Putnick & Bornstein, 2016) because they are willing to accept models that are imperfect but useful approximations of a population (MacCallum, 2003). Our current investigation focuses only on the χ^2_{MV} test statistic.[6]

If configural invariance appears tenable, then additional restrictions on model parameters can be tested. Metric (or "weak") invariance can be evaluated by constraining factor loadings to equality across groups and testing those constraints with the $\Delta\chi^2_{MV}$ statistic. If metric invariance holds, then scalar (or "strong") invariance can be tested by additionally constraining intercepts (of continuous indicators) to equality across groups, and/or strict invariance can be tested by additionally constraining residual variances to equality across groups; again, those additional constraints can be tested with the $\Delta\chi^2_{MV}$ statistic. With ordered-categorical indicators, testing metric, scalar, and strict invariance is not as straight-forward, due to additional identification requirements described next.

Because a latent variable is not directly observed, it has no scale of measurement. The common factors in CFA models are assumed to be normally distributed random variables, but the mean and variance of that distribution is not identified without additional constraints on estimated model parameters. Choices between identification constraints are essentially arbitrary in a single-group cross-sectional context, but they could limit the comparisons one could make across groups or occasions in multi-group or longitudinal models (Millsap & Yun-Tein, 2004; Muthén & Asparouhov, 2002). Incorporating a threshold model in IFA models further complicates the matter because the same identification issue holds for each ordered-categorical indicator's latent item-response. Each latent item-response is also assumed to be a normally distributed random variable whose mean and variance are not identified without additional constraints on estimated model parameters.

A simple identification method is to fix the common factor mean and variance to zero and one, respectively, so that common factors are assumed *standard* normally distributed. Alternative identification constraints include fixing a single factor loading per factor to one, or constraining the average of all loadings per factor to be one (Little, Slegers, & Card, 2006). Latent item-response distributions can be identified the same way, but they are endogenous, so the mean of the distribution is conditional on the common factor (i.e., it is an intercept). The intercepts are typically fixed to zero, and the variances are identified either by

[4] Details about how to use the DIFFTEST command can be found with Web Note 4 at http://www.statmodel.com/ .

[5] Jorgensen et al. (2017a) showed that the χ^2 test of overall model fit tests an overly restrictive null hypothesis because model configurations could be equivalent across populations even if the hypothesized model is not a perfectly accurate representation of it. This issue is discussed elsewhere in greater detail (Jorgensen, 2017; Jorgensen et al., 2017), but it is beyond the focus of the current study, which focuses on situations in which the χ^2_{MV} test fails even in the ideal circumstance that the hypothesized model is a perfect representation of the population(s).

[6] Jorgensen et al. (2017a) showed that permuting alternative fit indices also provides valid tests of hypotheses about measurement invariance.

fixing each total variance to one (consistent with the so-called *delta* method in M*plus* and lavaan; see Muthén & Asparouhov, 2002, for details) or by fixing each residual variance to one (as it would be in a probit regression model, called the *theta* method in M*plus* and lavaan). Because intercepts are typically fixed to zero for identification, their equality across groups is assumed rather than tested. In practice, invariance of thresholds is often substituted for invariance of intercepts when assessing scalar invariance (Putnick & Bornstein, 2016; see Millsap & Yun-Tein, 2004; for differences between M*plus* and LISREL, the latter of which assumes invariance of all thresholds and instead tests equality of intercepts), although it is worth noting that mean and covariance structures are not identified independent of each other in IFA as they are in CFA (Millsap & Yun-Tein, 2004; Muthén & Asparouhov, 2002; Wu & Estabrook, 2016). After applying equality constraints on factor loadings and thresholds, the theta identification method allows residual variances to be freely estimated in all but one group, leading Millsap and Yun-Tein (2004) to recommend theta over delta parameterization so that strict invariance could be explicitly tested. However, because strict invariance is not required to draw inferences about common-factor distributions (e.g., to compare latent means, variances, or correlations between groups), strict invariance is not tested as often as metric or scalar invariance (Putnick & Bornstein, 2016). For brevity, our investigation does not focus on strict invariance, and we use the delta method of identification employed by Sass et al. (2014).

It is important to note when identifying the common-factor means and variances by fixing them to zero and one, respectively, in the configural model (as described above), some identification constraints can be relaxed when equality constraints are placed upon measurement parameters (Kline, 2016; Little, 2013). For instance, when factor loadings are constrained to equality across groups in the metric model, the factor variance only needs to be constrained to one for a single group. Likewise, only one group's factor mean(s) need to be fixed to zero in the scalar model. If the common-factor means and variances are not freed when they are no longer required for identification, then the $\Delta\chi^2_{MV}$ test would not only test the null hypothesis of equal measurement parameters, but rather would simultaneously test null hypotheses of equal measurement parameters *and* common-factor distributions.

Similarly, when using the delta method to identify latent item-response scales, all latent item-response variances are constrained to equal one in the configural model by fixing the residual variances to one minus the variance explained by the common factor(s). When loadings and thresholds are constrained to equality across groups, latent item-response scales only need to be constrained to one in a single group, so they can be freely estimated in other groups (Wu & Estabrook, 2016). Thus, constraining a binary item's loading and single threshold to equality across groups only results in gaining one degree of freedom because its residual variance must be freed. This means metric and scalar invariance cannot be tested separately for binary items[7] (Millsap & Yun-Tein, 2004). Thus, like Sass et al. (2014), we focus only on a simultaneous test of the omnibus null hypothesis that both loadings and thresholds are equivalent across groups. This is a common method in applied research, as well (e.g., Garnaat & Norton, 2010; Randall & Engelhard, 2010).

PERMUTATION TEST OF MEASUREMENT INVARIANCE

The focus of the present research is to evaluate random permutation, a nonparametric method, when applied to tests of measurement invariance in IFA models. Interested readers can consult Hayes (1996), who provided an excellent introduction to permutation tests in general, and Rodgers (1999), who placed permutation in a taxonomy of other resampling procedures (e.g., bootstrapping). Jorgensen et al. (2017a) first proposed a permutation test of measurement invariance in CFA models with multivariate normally distributed indicators, and the same steps can be applied in the case of ordered-categorical indicators. The procedure involves the steps enumerated below.

When testing configural invariance, these steps are applied only to the configural model, and the χ^2 test statistic is saved in Steps 1 and 4. When testing scalar invariance, both the configural and scalar models are fitted in Steps 1 and 4, and the $\Delta\chi^2$ test statistic is saved. It is not necessary to calculate a robust $(\Delta)\chi^2_{MV}$ test statistic, but results would be equivalent (within reasonable Monte Carlo error; Jorgensen et al., 2017a) because the *p* value for either statistic would be based on the empirical permutation distribution of that statistic.

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

[7] Wu and Estabrook (2016) recently showed that it is not possible to test equality of thresholds independently of any other type of measurement parameter. It is only possible to test equality of thresholds on the condition of at least one other type of measurement parameter (for items with four or more categories), at least two other types (for items with three categories), or at least three other types (for binary items). This finding has implications for how measurement invariance should be tested with ordered-categorical indicators, but such a paradigm shift is beyond the scope of the current article.

5. Repeat Steps 2–4 I times, resulting in a vector of length I for each fit measure.
6. Make an inference about $(\Delta)\chi^2$ by comparing it to the vector of permuted $(\Delta)\chi^2$ statistics.

The number of times I that the grouping variable is randomly shuffled should be large enough to approximate the true sampling distribution. We used a preliminary study (see Appendix A[8]) to determine that $I = 500$ random shuffles is sufficient to minimize Monte Carlo error when estimating the p value in Step 6. Step 6 can be accomplished in either of two ways, yielding the same decision about the null hypothesis:

- Calculate the proportion of the vector of $(\Delta)\chi^2$ statistics that is more extreme (i.e., indicates worse fit or a greater decrement in fit) than the observed $(\Delta)\chi^2$. This is an approximate one-tailed p value that approximates the probability of obtaining a $(\Delta)\chi^2$ at least as poor as the observed one, if the null hypothesis of invariance across all groups holds true. Reject the null hypothesis if $p<\alpha$.
- Sort the vector of permuted $(\Delta)\chi^2$ statistics in ascending order. Use the $(100 \times (1-\alpha))$th percentile as a critical value, and reject H0 if $(\Delta)\chi^2$ is more extreme than the critical value.

Jorgensen et al. (2017a) showed that permutation provided nominal Type I error rates and sufficient power to detect substantial violations of invariance, regardless of the fit measure used as criterion (i.e., $(\Delta)\chi^2$ or an alternative fit index). This provides an additional potential advantage of permutation for researchers who prefer to assess invariance using (change in) indices of approximate fit, given that Sass et al. (2014) found the guidelines based on past simulations studies yield inflated Type I errors with ordered-categorical indicators. According to Putnick and Bornstein (2016), the most frequently consulted statistic to assess invariance in CFA is a ΔCFI cutoff (effectively, a critical value) of .01 (Cheung & Rensvold, 2002), although others are also frequently referenced, such as a Δ*RMSEA* cutoff of .01 (Chen, 2007). Jorgensen et al. (2017a) showed that Type I error inflation of fit indices is more likely with small samples, when their sampling distributions have greater variance, making it more likely that samples produce estimates below these proposed cutoffs. Thus, although our investigation focuses on the $(\Delta)\chi^2_{MV}$ statistic, readers should be aware that the results herein generalize to other measures of model fit.

[8] Appendix A also discusses the issue of sparse data, when not all levels of a variable are observed in each group.

TABLE 1
Fit Metrics from Empirical Example

Software:	*M*plus *and DIFFTEST*		*lavaan and lavTestLRT*		*Permutation*	
Model:	*Configural*	*Scalar*	*Configural*	*Scalar*	*Configural*	*Scalar*
χ^2	55.925	64.348	33.560	56.642	25.913	44.518
df	18	40	18	40	18	40
$p(\chi^2)$	< .001	.008	.014	.042	.340	.522
$\Delta\chi^2$		21.590		12.478		18.604
Δdf		22		22		22
$p(\Delta\chi^2)$		.485		.254		.354
RMSEA	0.103	0.055	0.066	0.046		
CFI	0.974	0.984	0.989	0.989		
TLI	0.957	0.988	0.982	0.991		

Note. The test statistics for the permutation tests are the unadjusted chi-squared values provided by lavaan, however the p-values were computed via permutation testing.

Empirical example

To compare the procedure of invariance testing in M*plus* and lavaan, we used the quality of life data analyzed by Chen and Yao (2015) through World Health Organization Quality-of-Life Scale (WHOQOL-BREF) as an illustrative example. For simplicity, in the current research, we only used the complete[9] cases (male: 158, female: 238) to test measurement invariance across gender of the psychological domain in WHOQOL-BREF (i.e., a single-factor model with six indicators measured on a 5-point Likert-type scale). The fit measures of configural and scalar invariance models are presented in Table 1. In our analysis the χ^2_{MV} statistics for the configural invariance model provided by M*plus* and lavaan were statistically significant; in practice, however, applied researchers using guidelines proposed by Hu and Bentler (1999) might consider the model fit indices to be sufficient for acceptance of the hypothesized model: CFI and TLI > 0.95 for both models in both software packages, although the configural model's RMSEA > 0.06 in both packages.

Regardless of the absolute fit of the model, the permutation test of configural invariance ($p = .340$) provided no evidence against the null hypothesis that men and women share equivalent population model configurations, whatever that true configuration might be.[10] Furthermore, the $\Delta\chi^2_{MV}$ statistics provided by M*plus* and lavaan when testing for scalar invariance were not statistically significant, nor was the permutation test of scalar invariance significant, indicating no evidence against the null hypothesis of scalar invariance.

[9] The application of the permutation method to incomplete data is a topic for future research that is beyond the scope of the current investigation.

[10] Jorgensen (2017) discussed modifying configurally invariant models with inadequate fit.

Hypotheses

The primary goal of our investigation is to compare the permutation test of invariance to the robust $(\Delta)\chi^2_{MV}$ test, which is the current gold standard, under conditions when $(\Delta)\chi^2_{MV}$ fails to adequately control the Type I error rate. Regarding tests of scalar invariance, Sass et al. (2014) found inflated error rates for five-category indicators when sample sizes were small to moderate (n = 150 or 300 per group). Such data represent a common situation in which researchers employ Likert-type response scales ranging, for example, from 1 (*Strongly Disagree*) to 5 (*Strongly Agree*). They observed this inflation regardless of whether thresholds were symmetric or asymmetric, but Bandalos (2014) and Rhemtulla et al. (2012) both found that asymmetry of thresholds can exacerbate inflated Type I error rates of the χ^2_{MV} statistic for testing overall model fit. Rhemtulla et al. (2012) also found that the number of categories influences the degree of inflation. Although Bandalos (2014) and Rhemtulla et al. (2012) did not investigate multi-group models, it can be expected that if these conditions hold for multiple groups, then χ^2_{MV} will also yield inflated errors in multi-group models. Because the χ^2_{MV} statistic is used to assess configural invariance, a permutation test could provide better control of Type I errors under these conditions. Chen (2007) found that the ratio of group sample sizes affected power to detect violations of invariance in CFA with multivariate normally distributed indicators. Because she did not investigate unbalanced groups when studying Type I error rates, we are also interested in investigating that possible effect in IFA with ordered-categorical indicators.

In line with past research, we hypothesize that for tests of configural invariance, χ^2_{MV} will yield inflated Type I error rates when items have asymmetric thresholds, that inflation will be more severe in small than in moderate samples, and that inflation will be more severe when items have fewer than more categories, but we expect permutation to yield nominal Type I error rates across all conditions. We hypothesize that $\Delta\chi^2_{MV}$ will yield inflated Type I error rates when testing scalar invariance in small samples, although there is no evidence that threshold asymmetry or number of categories will affect this inflation. We have no expectations about whether unbalanced sample sizes will affect Type I error rates for χ^2_{MV} or $\Delta\chi^2_{MV}$, but we expect them to be nominal using permutation. Because past research showed that Type I error inflation dissipated in very large samples (e.g., N = 1000), we have no need to compare the permutation method to $(\Delta)\chi^2_{MV}$ in larger samples.

Regarding software implementations, we expect the exact method for calculating the Jacobian of the constraint function to control Type I error rates at least as well as the approximate delta method, although we know of no previous research on which to base any more specific hypotheses. Because both methods are available using lavTestLRT, we use the lavaan package in our investigation. Because most other simulation research we reviewed (Bandalos, 2014; Chen, 2007; French & Finch, 2006; Kim & Yoon, 2011; Rhemtulla et al., 2012; Sass et al., 2014) used M*plus*, we also fit models using M*plus*.

Although we expect Type I error rates to be closer to nominal using permutation than $(\Delta)\chi^2_{MV}$ in conditions stated above, if the differences are not substantial, then it would be of practical interest to know whether better control of Type I errors comes at the expense of losing adequate power to detect true violations of invariance. To this end, we also explore power as an outcome of secondary interest.

METHOD

Sass et al. (2014) simultaneously tested metric and scalar invariance constraints while manipulating many of the factors whose effects we are interested in. We therefore partially replicated their Monte Carlo simulation so that we could compare our results to theirs. For factors that we manipulated that they did not, we borrowed conditions from other studies so that our results could be compared to previous research. Whereas Whereas Sass et al. (2014) investigated only tests of scalar invariance, we investigated tests of both configural and scalar invariance.

Data for two groups were generated in R using the simulateData function in the lavaan package, which simulates latent-item responses from multivariate normal distributions implied by specific population parameters, then applies specified population thresholds to discretize latent item-responses. Borrowing Sass et al.'s (2014) simulation conditions, we simulated a single latent common factor ($\mu = 0$ and $\sigma = 1$ in both groups) measured by eight indicator variables with factor loadings of $\lambda = .60$ and residual variances of $1 - \lambda^2 = .64$. Following Sass et al. (2014), we manipulated the values of thresholds, as well as the number of thresholds (i.e., response categories), total and group sample sizes, and for tests of scalar invariance, whether factor loadings were invariant. We simulated 1000 replications in each condition, described below.

Response categories

Sass et al. (2014) only simulated data with five response options per indicator, but we also simulated data with two response options because previous simulations conflicted about the degree to which results differ (Rhemtulla et al., 2012) or not (Bandalos, 2014) across this factor. We chose dichotomous and five-category polytomous data to mimic response scales that are often used in applied research. For example, binary data represent "Yes" or "No" responses indicating whether a statement is true. Data with five response options represent response scales where, for example, participants are asked to indicate their level of agreement with a

statement on a Likert-type scale from 1 (*Strongly Disagree*) to 5 (*Strongly Agree*). Only two levels were chosen because previous research has already investigated additional levels (Bandalos, 2014; Rhemtulla et al., 2012). Our goal of comparing Type I error rates across testing methods is served most parsimoniously by investigating lower and upper limits of the number of categories applied researchers would likely treat as categorical in practice. Rhemtulla et al. (2012) showed that robust maximum likelihood and DWLS produce similar enough results to treat ordered-categorical data as continuous when there are at least five response categories (except in some asymmetric-threshold conditions), so we use that as our upper-limit condition, and binary as our lower limit.

Threshold symmetry

We specified population thresholds (z scores from a standard normal distribution for each latent item-response) that yielded symmetric or asymmetric observed indicator distributions. Generating data using asymmetric thresholds was done in order increase generalizability to applied research where response distributions are often asymmetric (e.g., the Serious Harm Reduction Scale; Martens, Pederson, LaBrie, Ferrier, & Cimini, 2007) and to replicate previous simulation work (Bandalos, 2014; Rhemtulla et al., 2012; Sass et al., 2014). We simulated symmetric dichotomous distributions using a single threshold of 0, similar to previous research (Beauducel & Herzberg, 2006; Rhemtulla et al., 2012). We simulated symmetric five-category polytomous data using thresholds of -1.30, -0.47, 0.47, and 1.30, chosen by Sass et al. (2014) to simulate conditions that generalize to applied research.

For tests of configural invariance, we simulated asymmetric dichotomous data using a single threshold of 1.198, and asymmetric polytomous data using thresholds of 0.85, 1.10, 1.45, and 2.00, in order to compare our results to those of Bandalos (2014). Although she did not report the exact thresholds she specified, she reported that they yielded expected skew = 2.40 and kurtosis = 3.78, noting that "the terms skew and kurtosis are not germane to categorical data" (Bandalos, 2014, p. 105). These were the least extreme values of skew and kurtosis she simulated, but they were more extreme than values used in other simulation research (Rhemtulla et al., 2012; Sass et al., 2014).

For tests of scalar invariance, we simulated asymmetric dichotomous data using a threshold of 0.70, which we chose as a compromise between what Rhemtulla et al. (2012) defined as moderate (0.36) and extreme (1.04) asymmetry. We simulated asymmetric polytomous data using thresholds of −0.25, 0.38, 0.84, and 1.28, which Sass et al. (2014) indicated were representative of "those typically found in practice" (p. 170).

Sample sizes and ratios

Sass et al. (2014) simulated total sample sizes of N = 300, 600, or 1000, with equal sizes in each of two groups. These are typical values in simulation research (Elosua, 2011; Elosua & Wells, 2013; Lubke & Muthén, 2004; Sass et al., 2014), but because Bandalos (2014) and Sass et al. (2014) only found inflated Type I errors at small and moderate sample sizes, we only simulated total sample sizes of N = 300 or 600. In a follow-up study, we also simulated one condition with N = 120 to verify whether permutation was still robust with very small sample sizes.

Sass et al. (2014) cited that Chen (2007) had found group ratios influenced power to detect non-invariance, but did not manipulate that factor themselves. Because Chen (2007) did not investigate whether group ratios effected Type I error rates, we varied group ratios to be either equal across groups (1:1 ratio) or to have twice as many cases (1:2 ratio) in the focal group (i.e., the group receiving the non-invariance manipulation). Holding total sample size constant at the levels indicated above, this resulted in group sample sizes of n = 150 and 150 or n = 100 and 200 in conditions with N = 300, or in group sample sizes of n = 300 and 300 or n = 200 and 400 in conditions with N = 600.

Factor loading invariance

Our main goal was to investigate whether permutation could provide better control of Type I error rates for tests of configural and scalar invariance than $(\Delta)\chi^2_{MV}$, which previous research has shown yields inflated errors under common conditions of small to moderate sample sizes and asymmetric thresholds. Therefore, we simulated conditions of full measurement invariance. However, better control of Type I errors is expected to come at the expense of lower power, so we also investigated power to detect a violation of invariance. We therefore simulated factor loadings in the focal group (Group 2) to be 0.35 rather than 0.60 for Items 1 and 2. This factor-loading difference of 0.25 was used by Sass et al. (2014) as well as French and Finch (2006), and is between the small (0.15) and large (0.40) conditions simulated by Stark, Chernyshenko, and Drasgow (2006).

We did not manipulate additional levels of factor-loading differences, nor did we manipulate other factors known to affect power (e.g., number or proportion of non-invariant items) because we were not interested in reiterating previous results (Chen, 2007; French & Finch, 2006; Kim & Yoon, 2011; Sass et al., 2014; Stark et al., 2006). Furthermore, Sass et al. (2014) showed that power to detect violations of invariance were similar for loadings and thresholds. So simulating one level of violated invariance was sufficient for our stated purpose of comparing power across different test procedures.

Models and test statistics

Tests of configural invariance were conducted using the χ^2_{MV} statistic available in lavaan by requesting estimator = "WLSMV", which implies estimator = "DWLS"

and test = "scaled.shifted". We also recorded the χ^2_{MV} statistic available in M*plus* by requesting "ESTIMATOR = WLSMV", which also implies DWLS estimation and a mean- and variance-adjusted test statistic. For comparison purposes, we also recorded the results of the unadjusted χ^2 test provided by lavaan. We conducted the permutation test using the unadjusted χ^2 statistic provided by lavaan. In all permutation tests, we used 500 shuffles of the grouping variable (see Appendix A for details).

The configural model was identified by fixing the factor mean and variance to zero and one, respectively, in both groups, and using the delta method (i.e., fixing residual variances of latent item-responses to $1 - \hat{\lambda}^2$, so that the total variance of each latent item-response equals one). All factor loadings and thresholds were freely estimated in both groups. The scalar model constrained estimated factor loadings and thresholds to be equivalent across groups, and the common-factor mean and variance were freely estimated in Group 2. In models with dichotomous items, $\Delta df = 14$, whereas with polytomous items, $\Delta df = 38$.[11]

Tests of scalar invariance were conducted using the $\Delta\chi^2_{MV}$ statistic available in M*plus* using the DIFFTEST procedure, and from lavaan using the lavTestLRT function with the argument method = "Satorra (2000)", which we requested using both the A.method = "exact" and the approximate A.method = "delta" methods for calculating the Jacobian of the constraint function. For comparison purposes, we also recorded the results of the unadjusted $\Delta\chi^2$ test using the unadjusted χ^2 values provided by lavaan. We also conducted the permutation test using this unadjusted $\Delta\chi^2$ statistic.

The outcome of interest was the proportion of replications in which the null hypothesis of invariance was rejected using each test, using $\alpha = .05$ or .01 as criterion for significance. When the populations were invariant, the null hypotheses were true, so rejection rates should be nominal, within reasonable Monte Carlo sampling error. Using a normal approximation to the binomial distribution given 1000 trials (replications), a 95% CI around a nominal value of .05 indicates a nominal range of .0365 to .0635 (Sass et al., 2014), and a nominal range of .0038 to .0162 using $\alpha = .01$. Test statistics that showed Type I errors within these nominal ranges would be considered to have acceptable Type I error control. In conditions in which measurement invariance was violated, the null hypothesis was false, so the rejection rate is an estimate of power. Methods that showed acceptable Type I error rates were compared on power. Specifically, determining whether the random permutation test controls Type I errors and yields power similar to or greater than other testing approaches was the main focus when interpreting the results.

[11] If we had fixed the factor means and variances in both groups even in the scalar model, as Sass et al. (2014) did, these differences would have been $\Delta df = 16$ and 40, respectively, as Sass et al. (2014) reported. We discuss the implications of this difference in the Discussion section.

RESULTS

The exact method for calculating the Jacobian of the constraint function yielded Type I error rates near zero and very low power in a subset of study conditions (see Appendix B), so we focus only on presenting the results of the approximate delta method, the M*plus* results, and the permutation results, along with unadjusted $(\Delta)\chi^2$ results from lavaan for comparison to a nonrobust method. Recall that the exact method was the default in previous versions of lavaan, but the approximate delta method is the default since version 0.6–1.

Configural invariance

The results of the configural invariance simulation showed that the random permutation test had acceptable Type I errors within the nominal ranges in all but one condition. The mean- and variance-adjusted χ^2 tests provided by M*plus* and lavaan performed nearly identically, and showed inflated Type I errors in conditions with asymmetric thresholds with five response categories. The error rates were especially inflated with five response categories when the total sample size was 300 (25.3%, 24.0%, 20.2%, and 19.9%). Lastly, the unadjusted χ^2 test provided by lavaan showed error rates well below the nominal range in all conditions (see Table 2), which is expected given that the unadjusted χ^2 statistic is calculated from a discrepancy function between model-implied polychoric correlations and the correlations estimated in a previous step, without taking into account the uncertainty associated with those estimated correlations.

Scalar invariance

Results showed that random permutation testing had reasonable Type I error control in all 16 conditions with equal measurement parameters, whereas M*plus* DIFFTEST and lavTestLRT had reasonable Type I error control in most of those conditions. The M*plus* DIFFTEST procedure had inflated error rates in five conditions, and the lavTestLRT procedure with the approximate Jacobian showed an inflated Type I error rate only in the condition with two response options, asymmetric thresholds, and 300 cases in each group. The majority of conditions with unbalanced samples had errors below the nominal range using lavTestLRT. The unadjusted χ^2 showed inflated Type I errors (see Table 3). The observed Type I error rates of the random permutation test, M*plus* DIFFTEST, and lavTestLRT were close enough to the nominal range (e.g., within .03 above or below the nominal .05) to warrant comparing their power to detect true violations of invariance.

Table 4 shows that the random permutation test, M*plus* DIFFTEST, and lavTestLRT have similar power in each of the 16 conditions. Among these three tests, the M*plus* DIFFTEST procedure consistently showed the highest power (also the most inflated Type I error rates), with

TABLE 2
Configural Invariance Type I Errors

Condition Info	*Permutations*	*M*plus	*lavaan*	*Unadjusted*
1:1, 300, 2, Symmetric	.051(.007)	.049(.014)	.049(.014)	**.001(.000)**
1:1, 600, 2, Symmetric	.048(.012)	.052(.012)	.052(.012)	**.001(.000)**
1:1, 300, 5, Symmetric	.042(.008)	.066(.009)	**.066**(.009)	**.000(.000)**
1:1, 600, 5, Symmetric	.041(.009)	.053(.012)	.053(.012)	**.000(.000)**
1:1, 300, 2, Non-Sym.	.046(.006)	.046(.008)	.046(.008)	**.004(.002)**
1:1, 600, 2, Non-Sym.	.052(.010)	.047(.008)	.047(.008)	**.000(.000)**
1:1, 300, 5, Non-Sym.	.057(.015)	**.200(.076)**	**.199(.075)**	**.022(.008)**
1:1, 600, 5, Non-Sym.	**.033**(.009)	**.090**(.012)	**.090**(.012)	**.001(.000)**
1:2, 300, 2, Symmetric	.041(.009)	.046(.008)	.047(.008)	**.000(.000)**
1:2, 600, 2, Symmetric	.051(.009)	.055(.008)	.055(.009)	**.000(.000)**
1:2, 300, 5, Symmetric	.042(.006)	.058(.010)	.058(.010)	**.000(.000)**
1:2, 600, 5, Symmetric	.048(.008)	.053(.010)	.053(.010)	**.000(.000)**
1:2, 300, 2, Non-Sym.	.047(.009)	.050(.007)	.050(.007)	**.005(.000)**
1:2, 600, 2, Non-Sym.	.056(.012)	.052(.008)	.052(.008)	**.001(.000)**
1:2, 300, 5, Non-Sym.	.055**(.017)**	**.253(.105)**	**.240(.088)**	**.035(.013)**
1:2, 600, 5, Non-Sym.	.051(.012)	**.106(.026)**	**.106(.027)**	**.002(.000)**

Note. Error rates are provided for α= .05 and α= .01 (parentheses). Values in bold fall outside of the nominal Type I error range.

TABLE 3
Scalar Invariance Type I Errors

Condition Info	*Permutations*	*M*plus *DIFFTEST*	*lavaan lavTestLRT*	*Unadjusted*
1:1, 300, 2, Symmetric	.046(.008)	.060(.020)	.056(.012)	**.143(.058)**
1:1, 600, 2, Symmetric	.040(.007)	.052(.012)	.050(.011)	**.128(.053)**
1:1, 300, 5, Symmetric	.046(.011)	.062(.015)	.054(.011)	**.131(.048)**
1:1, 600, 5, Symmetric	.051(.007)	.057(.008)	.053(.005)	**.098(.050)**
1:1, 300, 2, Non-Sym	.051**(.017)**	**.065(.020)**	.054(.014)	**.135(.053)**
1:1, 600, 2, Non-Sym	.053(.015)	**.078(.020)**	**.065(.017)**	**.139(.065)**
1:1, 300, 5, Non-Sym	.047(.009)	.053(.013)	.047(.008)	**.131(.050)**
1:1, 600, 5, Non-Sym	.054(.009)	.062(.013)	.056(.012)	**.128(.061)**
1:2, 300, 2, Symmetric	.049(.007)	**.069(.019)**	**.035(.002)**	**.157(.058)**
1:2, 600, 2, Symmetric	.061(.009)	**.071(.017)**	.041**(.003)**	**.154(.064)**
1:2, 300, 5, Symmetric	.039(.007)	.051(.011)	**.017(.003)**	**.099(.033)**
1:2, 600, 5, Symmetric	.056(.008)	.061(.011)	**.019(.002)**	**.119(.049)**
1:2, 300, 2, Non-Sym	.049(.010)	**.081(.021)**	**.030**(.006)	**.147(.059)**
1:2, 600, 2, Non-Sym	.054(.007)	**.072**(.012)	**.034(.002)**	**.131(.057)**
1:2, 300, 5, Non-Sym	.041(.008)	.051(.016)	**.022(.002)**	**.115(.038)**
1:2, 600, 5, Non-Sym	.050(.009)	.057(.012)	**.022(.002)**	**.125(.051)**

Note. Error rates are provided for α= .05 and α= .01 (parentheses). Values in bold fall outside of the nominal Type I error range.

lavTestLRT showing power equal to or greater than the random permutation test. However, differences in power between these three tests were typically negligible (e.g., < 4% difference between the highest and lowest power). All testing procedures showed higher power in conditions with higher total *N*, more response categories, and symmetric (vs. asymmetric) thresholds. Power was not always consistently higher or lower with (un)balanced groups, but holding other factors constant, unbalanced samples yielded slightly higher power than balanced samples when $N = 300$ than when $N = 600$, at least for random permutation and M*plus* DIFFTEST. But the power advantage typically occurred in conditions that also yielded inflated Type I errors.

When choosing a test, researchers must balance the costs of Type I and Type II errors (the complement of power). The comparisons between the different testing approaches on their ability to balance power and Type I error control is shown in Figure 1. Each plotted point represents a rejection rate when the null hypothesis of equal factor loadings was true (*x*-axis) and false (*y*-axis), for a given sample size (ratio), number of response categories, and threshold distribution. This figure provides a depiction of the balance between Type I error control and power of these testing approaches. Figure 1 demonstrates that

TABLE 4
Scalar Invariance Power

Condition Info	*Permutations*	*M*plus *DIFFTEST*	*lavaan lavTestLRT*	*Unadjusted*
1:1, 300, 2, Symmetric	.274(.108)	.319(.154)	.292(.121)	.452(.290)
1:1, 600, 2, Symmetric	.538(.300)	.568(.338)	.543(.313)	.703(.556)
1:1, 300, 5, Symmetric	.451(.228)	.504(.288)	.464(.241)	.618(.447)
1:1, 600, 5, Symmetric	.781(.582)	.811(.644)	.794(.604)	.890(.795)
1:1, 300, 2, Non-Sym	.207(.077)	.258(.110)	.225(.083)	.361(.214)
1:1, 600, 2, Non-Sym	.401(.202)	.456(.244)	.427(.205)	.588(.413)
1:1, 300, 5, Non-Sym	.337(.148)	.370(.197)	.335(.147)	.519(.338)
1:1, 600, 5, Non-Sym	.703(.477)	.733(.532)	.712(.490)	.831(.723)
1:2, 300, 2, Symmetric	.259(.105)	.319(.154)	.244(.102)	.449(.282)
1:2, 600, 2, Symmetric	.548(.319)	.585(.372)	.503(.266)	.702(.557)
1:2, 300, 5, Symmetric	.427(.214)	.459(.265)	.354(.139)	.566(.410)
1:2, 600, 5, Symmetric	.803(.611)	.822(.663)	.739(.506)	.885(.805)
1:2, 300, 2, Non-Sym	.215(.070)	.276(.121)	.208(.061)	.368(.229)
1:2, 600, 2, Non-Sym	.403(.197)	.436(.238)	.356(.158)	.583(.402)
1:2, 300, 5, Non-Sym	.341(.159)	.384(.207)	.274(.100)	.494(.332)
1:2, 600, 5, Non-Sym	.696(.468)	.731(.524)	.618(.343)	.827(.702)

Note. Rejection rates are provided for α = .05 and α = .01 (parentheses).

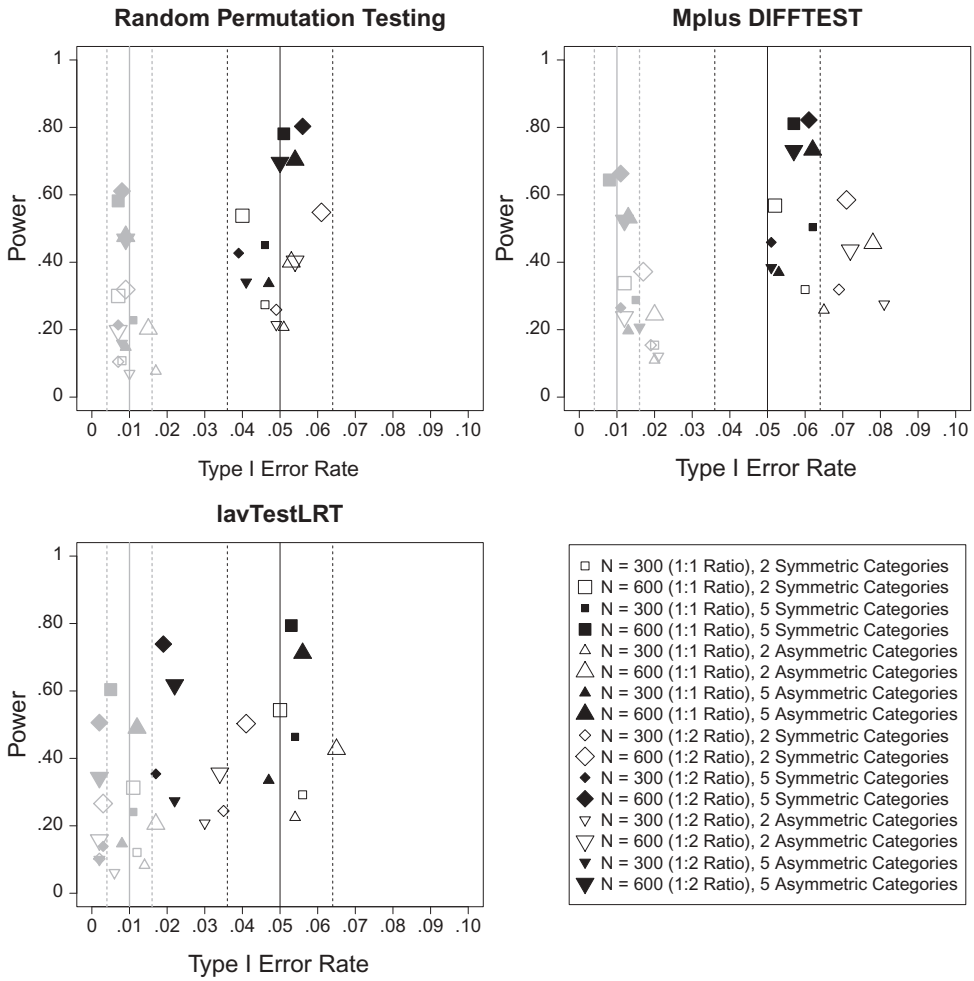

FIGURE 1 Type I errors and power across conditions. Squares indicate symmetric and triangles asymmetric thresholds. Outlines indicate binary data, and solid symbols indicate 5 categories. Larger symbols indicate larger sample sizes, and inverted symbols indicate unbalanced samples. Black represents results using α = .05, and grey represents α = .01. Solid lines for nominal α levels are surrounded by dotted lines indicating differences attributable to sample error (with 95% confidence).

the range of power was similar across test procedures (panels), yet the random permutation test most consistently controlled Type I errors across conditions. The cost of slightly higher power with the M*plus* DIFFTEST procedure comes at the cost of inflated Type I errors.

Small sample size follow-up

We conducted a brief follow-up simulation by generating measurement invariant binary data for two groups with 40 and 80 cases. This was done to evaluate the generalizability of the previously presented simulation results to cases with fewer observations. All model specifications were consistent with the previously described simulations; the population threshold value was set to 0, and all population factor loadings were set to .60. We attempted to simulate conditions with five categories and with asymmetric thresholds, but those conditions greatly increased the occurence for some categories not to be observed at all in the simulated data, as well as after permuting the grouping variable. We think an investigation of small sample size across more conditions would be ideal for comparing different methods of dealing with sparse data, as discussed in Appendix A, but that is beyond the scope of our study, so we suggest it for future research.

When testing configural invariance, the random permutation test (.046), M*plus* (.055), and lavaan (.056) showed acceptable error control with $\alpha = .05$; the same was true with $\alpha = .01$ with error rates of .014, .011, and .012, respectively. When testing for scalar invariance with $\alpha = .05$, the random permutation test (.052) showed superior error control compared to DIFFTEST procedure (.122) and lavTestLRT procedure (.068). However, with $\alpha = .01$ the random permutation test (.009) and the lavTestLRT procedure (.009) both performed well, whereas the error rate with DIFFTEST was inflated (.033).

DISCUSSION

Research questions

The purpose of the present research was to evaluate the use of random permutation testing applied to tests of measurement invariance with ordered-categorical indicator variables. When models with ordered-categorical data are estimated with DWLS, the χ^2 model-fit statistic requires a robust correction for tests to yield nominal Type I error rates. Recent simulation studies (Bandalos, 2014; Sass et al., 2014) have shown that large sample sizes are required for nominal error rates, especially when thresholds are asymmetric. We proposed permutation as an alternative robust test that is easily implemented in any statistical software without the need for a scaling correction, and as a method that can control Type I errors as well or better than the standard $(\Delta)\chi^2_{MV}$, particularly under conditions when $(\Delta)\chi^2_{MV}$ yields inflated error rates. The Monte Carlo simulation expanded on the work of Jorgensen et al. (2017a, 2017b) and served as a follow-up to recent studies by Bandalos (2014) and Sass et al. (2014).

Overall, the random permutation test performed well when testing both configural and scalar invariance. For configural models, we observed similar Type I error inflation from χ^2_{MV} as Bandalos (2014) observed when data were asymmetric; however, we only observed inflation for five-category data, not for binary data. Bandalos (2014), in contrast, did not report Type I error rates separately by number of categories, stating that "the effect of number of categories was not as pronounced" (p. 109). The random permutation yielded nominal Type I error rates across conditions, indicating the permutation test should be considered a viable alternative to χ^2_{MV} when evaluating configural invariance with small to moderate sample sizes. Given that configural models frequently do not fit the data perfectly (MacCallum, 2003; Putnick & Bornstein, 2016), the permutation test of configural invariance can also prevent inflated Type I errors when the model fits only approximately well (Jorgensen et al., 2017).

The random permutation test was the only method that consistently showed Type I errors within the previously defined nominal ranges across conditions. As expected, the power of the random permutation test was increased in conditions with larger group sizes, but also increased with more response categories and was greater with symmetric response distributions. As would be expected based on the better error control, the random permutation test showed slightly less power than M*plus* DIFFTEST and lavTestLRT.

The inflated error rates found by Sass et al. (2014) suggested that the error control of the M*plus* DIFFTEST procedure is poor; however, their findings could be attributable to the model identification conditions used in their simulation. Sass et al. (2014) compared two groups on 10 indicator variables with five response options, and in their configural model the common-factor means and variances were constrained to 0 and 1 in both groups for identification. However, they did not free those constraints in the second group after the measurement parameters were constrained to equality in their scalar model (Kline, 2016; Little, 2013). Their scalar model was therefore too restrictive to merely test for equality of measurement parameters because it implied that all model parameters were equal across groups. The less stringent, appropriate scalar-invariance model in our simulation resulted in Type I errors from the M*plus* DIFFTEST procedure that were closer to nominal levels than was reported by Sass et al. (2014).

Suggestions for random permutation testing

The results of the present research clearly show how the $(\Delta)\chi^2$ testing approaches compare on their ability to control Type I errors. Given the only slight differences in power across testing

procedures, the present research suggests that when researchers desire the Type I error rate of their test to be as close as possible to their α level, the random permutation testing procedure could be preferable to the parametric approaches evaluated here, at least in non-ideal situations (e.g., small or moderate unbalanced samples with asymmetric thresholds). With sufficiently large, balanced samples and symmetric thresholds, there is no reason to prefer permutation over the current gold standard, $(\Delta)\chi^2_{MV}$, unless the mean-and-variance adjustment cannot be calculated because the required W matrix cannot be inverted for a particular sample.

Limitations

The present research might not generalize to applied research situations because invariant models were exactly correctly specified, but Jorgensen et al. (2017a) showed that with more realistic imperfect models (MacCallum, 2003; Putnick & Bornstein, 2016), permutation controls Type I error rates better than the standard χ^2 test. Similarly, although unequal observed response categories across groups in permutation shuffles were infrequent in the present research, this might occur more frequently in applied research, particularly in very small samples. We employed a reshuffling method to deal with sparseness in our current investigation, and although our results indicated that p values could be trusted to make decisions, other options could be explored in future research designed specifically to test boundary conditions under which one or more alternative solutions could be expected to fail.

Our exploratory investigation of power only involved violating metric invariance in one condition, but previous research has shown additional factors influence power to detect violations of invariance, such as the number of factors and items per factor (Chen, 2007; French & Finch, 2006), the number (or proportion) of non-invariant items (Chen, 2007), and the type (e.g., metric or scalar; Kim & Yoon, 2011; Stark et al., 2006) and magnitude (French & Finch, 2006; Stark et al., 2006) of the violation. We were primarily interested in assessing Type I error control, and because there is no evidence that these factors affect that, we did not investigate these factors here. However, previous results suggest that similar power can be expected between the permutation test and $(\Delta)\chi^2_{MV}$ across a variety of conditions (Jorgensen, 2017; Jorgensen et al., 2017a).

Hayes (1996) showed that the permutation test is not entirely nonparametric. Its distributional assumptions are not as strict as parametric tests, but it does implicitly assume that observations are *exchangeable*. When subjects are randomly assigned to groups, the exchangeability assumption is met by design. In observational studies, this assumption can be violated to whatever degree the population distributions differ in (for example) variance. It remains to be investigated to what degree the method described herein would be robust to a violation of the exchangeability assumption.

Directions for future research

Clearly, additional Monte Carlo simulation research is needed to further evaluate the performance of the random permutation test. The present research was designed only to determine which variables influenced Type I error rates in boundary conditions of factors that have been shown to inflate errors in past research; future research is required to further probe those effects and offer suggestions. The random permutation test performance was overall quite similar to the M*plus* DIFFTEST method and lavTestLRT. The benefit of the random permutation test is that it can be implemented using any quantity of interest, so future research could evaluate random permutation testing using popular fit measures (e.g., RMSEA), particularly without known sampling distribution (e.g., CFI, TLI, or SRMR).

The strength of the permutation method lies in its flexibility. For example, equality of thresholds across groups can be tested one pair of items at a time without fitting a common-factor model, requiring only the assumption of bivariate normality for each pair of items (Jöreskog, 2002; Verdam, Oort, & Sprangers, 2016). Permutation methods have been developed for testing equivalence of between-group heterogeneity of ordinal variables (Arboretti Giancristofaro et al., 2009; Bonnini, 2014), which could be applicable to testing between-group equivalence of thresholds.

Conclusion

The present research provided a promising initial evaluation of random permutation testing to handle $(\Delta)\chi^2$ testing with ordered-categorical indicator variables. The permutation test provided nominal Type I errors under more conditions than the parametric testing approaches that represent the current gold standard. The present research suggests that researchers should consider the random permutation testing procedure a viable option for $(\Delta)\chi^2$ tests of measurement invariance with ordered-categorical indicator variables.

ACKNOWLEDGMENT

We would like to thank Yves Rosseel for his helpful technical discussions while investigating different implementations of the mean- and variance-adjusted test statistic, and Paul Johnson for his computational assistance while comparing software packages. We thank the Center for Research Methods and Data Analysis and the College of Liberal Sciences at the University of Kansas for access to their high performance compute cluster on which our Monte Carlo simulations were conducted.

REFERENCES

Arboretti Giancristofaro, R., Bonnini, S., & Pesarin, F. (2009). A permutation approach for testing heterogeneity in two-sample categorical variables. *Statistics and Computing*, *19*(2), 209–216. doi:10.1007/s11222-008-9085-8

Asparouhov, T., & Muthén, B. (2006). *Robust chi square difference testing with mean and variance adjusted test statistics* (Technical report). Muthén and Muthén. Mplus Web Notes: No. 10.

Bandalos, D. L. (2014). Relative performance of categorical diagonally weighted least squares and robust maximum likelihood estimation. *Structural Equation Modeling*, *21*(1), 102–116. doi:10.1080/10705511.2014.859510

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in cfa. *Structural Equation Modeling*, *13*(2), 186–203. doi:10.1207/s15328007sem1302_2

Bentler, P. M., & Satorra, A. (2010). Testing model nesting and equivalence. *Psychological Methods*, *15*(2), 111–123. doi:10.1037/a0019625

Bonnini, S. (2014). Testing for heterogeneity with categorical data: Permutation solution vs. bootstrap method. *Communications in Statistics-Theory and Methods*, *43*(4), 906–917. doi:10.1080/03610926.2013.799376

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*(3), 464–504. doi:10.1080/10705510701301834

Chen, P.-Y., & Yao, G. (2015). Measuring quality of life with fuzzy numbers: In the perspectives of reliability, validity, measurement invariance, and feasibility. *Quality of Life Research*, *24*, 781–785. doi:10.1007/s11136-014-0816-3

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233–255. doi:10.1207/S15328007SEM0902_5

Elosua, P. (2011). Assessing measurement equivalence in ordered-categorical data. *Psicologica*, *32*(2), 403–421.

Elosua, P., & Wells, C. S. (2013). Detecting dif in polytomous items using macs, irt and ordinal logistic regression. *Psicologica*, *34*(2), 327–342.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466–491. doi:10.1037/1082-989X.9.4.466

French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(3), 378–402. doi:10.1207/s15328007sem1303_3

Garnaat, S. L., & Norton, P. J. (2010). Factor structure and measurement invariance of the yale-brown obsessive compulsive scale across four racial/ethnic groups. *Journal of Anxiety Disorders*, *24*(7), 723–728. doi:10.1016/j.janxdis.2010.05.004

Hayes, A. F. (1996). Permutation test is not distribution-free: Testing H_0: $\rho = 0$. *Psychological Methods*, *1*(2), 184–198. doi:10.1037/1082-989X.1.2.184

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), 1–55. doi:10.1080/10705519909540118

Jöreskog, K. G. (2002). *Structural equation modeling with ordinal variables using LISREL*. Retrieved from http://www.ssicentral.com/lisrel/techdocs/ordinal.pdf.

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research*, *36*(3), 347–387. doi:10.1207/S15327906347-387

Jorgensen, T. D. (2017). Applying permutation tests and multivariate modification indices to configurally invariant models that need respecification. *Frontiers in Psychology*, *8*, 1455. doi:10.3389/fpsyg.2017.01455

Jorgensen, T. D., Kite, B., Chen, P.-Y., & Short, S. D. (2017a). Permutation randomization methods for testing measurement equivalence and detecting differential item functioning in multiple-group confirmatory factor analysis. *Psychological Methods*. Advanced online publication. doi: 10.1037/met0000152

Jorgensen, T. D., Kite, B., Chen, P.-Y., & Short, S. D. (2017b). 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, North Carolina, 2016* (pp. 93–103). New York, NY: Springer.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling*, *15*(1), 136–153. doi:10.1080/10705510701758406

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical cfa and irt. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(2), 212–228. doi:10.1080/10705511.2011.557337

Kline, R. B. (2016). *Principles and Practice of Structural Equation Modeling* (4 ed.). New York, NY: Guilford.

Little, T. D. (2013). *Longitudinal Structural Equation Modeling*. New York, NY: Guilford Press.

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: A Multidisciplinary Journal*, *13*(1), 59–72. doi:10.1207/s15328007sem1301_3

Lubke, G. H., & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to likert scale data complicates meaningful group comparisons. *Structural Equation Modeling*, *11*(4), 514–534. doi:10.1207/s15328007sem1104_2

MacCallum, R. C. (2003). 2001 presidential address: Working with imperfect models. *Multivariate Behavioral Research*, *38*(1), 113–139. doi:10.1207/S15327906MBR3801_5

Martens, M. P., Pederson, E. R., LaBrie, J. W., Ferrier, A. G., & Cimini, M. D. (2007). Measuring alcohol-related protective behavioral strategies among college students: Further examination of the protective behavioral strategies scale. *Psychology of Addictive Behaviors*, *21*(3), 307–315. doi:10.1037/0893-164X.21.3.307

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*(3), 479–515. doi:10.1207/S15327906MBR3903_4

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132. doi:10.1007/BF02294210

Muthén, B., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus* (Technical report). Muthén and Muthén. Mplus Web Notes: No. 4.

Muthén, L. K., & Muthén, B. O. (1998–2015). *Mplus user's guide* (7 ed.). Los Angeles, CA: Muthén and Muthén.

Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kickpatrick, R. M., … Boker, S. M. (2016). OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika*, *81*(2), 535–549. doi:10.1007/s11336-014-9435-8

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. doi:10.1016/j.dr.2016.06.004

R Core Team. (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrived from https://www.R-project.org/

Randall, J., & Engelhard, G. (2010). Using confirmatory factor analysis and the rasch model to assess measurement invariance in a high stakes reading assessment. *Applied Measurement in Education*, *23*(3), 286–306. doi:10.1080/08957347.2010.486289

Rhemtulla, M., Brosseau-Liard, P. E., & Savalei, V. (2012). When can categorical variables be treated as continuous? a comparison of robust continuous and categorical sem estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. doi:10.1037/a0029315

Rodgers, J. L. (1999). The bootstrap, the jackknife, and the randomization test: A sampling taxonomy. *Multivariate Behavioral Research*, *34*(4), 441–456. doi:10.1207/S15327906MBR3404_2

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. doi: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*(2), 167–180. doi:10.1080/10705511.2014.882658

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In R. D. H. Heijmans, D. S. G. Pollock, & A. Satorra (Eds.), *Innovations in multivariate statistical analysis: A Festschrift for Heinz Neudecker* (pp. 233–247). London, England: Kluwer Academic Publishers.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, *91*(6), 1292. doi:10.1037/0021-9010.91.6.1292

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*(1), 4–70. doi:10.1177/109442810031002

Verdam, M. G. E., Oort, F. J., & Sprangers, M. A. G. (2016). Using structural equation modeling to detect response shifts and true change in discrete variables: An application to the items of the sf-36. *Quality of Life Research*, *25*(6), 1361–1383. doi:10.1007/s11136-015-1195-0

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*(1), 58–79. doi:10.1037/1082-989X.12.1.58

Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika*, *81*(4), 1014–1045. doi:10.1007/s11336-016-9506-0

APPENDIX A

Prior to conducting the larger Monte Carlo simulations, we evaluated the influence of the number of group shuffles on the random permutation test results. A subset of conditions used in the main simulation was used in this pilot study. The performance of the test was evaluated in data-generation conditions with different numbers of response options (2 vs. 5) and manipulated factor loading invariance (equal in all groups vs. different for the first two items). Sample sizes were held constant at $N = 150$ per group, using the symmetric thresholds specified in the main study. In each replication the random permutation test was conducted with 100 through 1,000 shuffles in increments of 100. The outcome of interest was the proportion of replications in which the addition of 100 permutation shuffles changed the rejection decision (i.e., reject or fail to reject the null hypothesis).

The results in Table A1 show the proportion of replications in which the rejection decision changed with an increase in the number of permutation shuffles. For example, when the number of permutation shuffles increased from 100 to 200, the decision about the null hypothesis of measurement invariance changed in 3.3% of the conditions with five response options and population invariance. Based on these results it was determined that beyond 500 shuffles there were few changes in the results. Beyond 500 shuffles, there was little change in rejection decisions, with less than 1% change in all four conditions. Therefore, 500 was determined to be a sufficient number for the Monte Carlo simulations in the present research.

TABLE A1
Proportion of Conditions with Change in Rejection Decision

	Two Response Options		*Five Response Options*	
Increase	*Invariance*	*Non Invariance*	*Invariance*	*Non Invariance*
100 to 200	.010	.033	.007	.028
200 to 300	.006	.018	.005	.016
300 to 400	.001	.014	.006	.015
400 to 500	.001	.013	.006	.008
500 to 600	.000	.010	.003	.009
600 to 700	.001	.008	.000	.008
700 to 800	.000	.011	.001	.005
800 to 900	.000	.005	.000	.004
900 to 1000	.000	.004	.001	.003

Note. Rejection rates are provided for $\alpha = .05$.

An important yet easily overlooked detail about using the permutation approach with ordered-categorical indicators is that the number of observations per response category within each group limits the number of possible permutations for which all categories contain at least one observation in each group after shuffling the grouping variable. This has practical consequences on the number of random shuffles that yield results in the empirical permutation distribution, perhaps requiring more than 500 shuffles to obtain 500 observed statistics. Consider, for example, the situation that only one (or a few) respondent(s) in a single group endorsed a particular response option for a discrete indicator. In such a case, there would be fewer thresholds for the same indicator in one group than another, and SEM software such as M*plus* would return an error because the same parameters cannot be estimated in all groups. One solution to which researchers might have to resort would be collapsing categories in the observed data prior to fitting their model. Although this ignores potentially useful information about individual differences, there is not enough such information in all groups to fit the same model to all groups.[12]

In circumstances where a single indicator variable has very few responses on an extreme end of a response scale for all groups, it is more likely that a random shuffle would result in all of those observations belonging to one group in a permuted data set. This could be handled in a few possible ways:

- Before shuffling the grouping variable, the rows of data susceptible to shuffling could be restricted to those without responses in the sparse categories. That is, the observations with responses in sparse categories could remain in their original groups, and all other observations could be randomly shuffled. One disadvantage of this constraint is that it would likely become prohibitively difficult if more than one indicator had a sparse category. Another disadvantage is that the restricted permutation distribution might not be a random sample from the full permutation distribution, yielding biased estimates of *p* values.

[12] If software is flexible enough (e.g., general Bayesian modeling software, or more flexible SEM software like OpenMx), it is possible to fit a model to each group that estimates only the thresholds between categories that were observed within each group. Equality constraints could still be imposed on loadings and the thresholds the researcher knows correspond to categories on the same response scale used in each group.

- A simpler solution would be to randomly shuffle the grouping variable again whenever the number of observed categories is not equal across groups. This also has the same potential disadvantage that the resulting distribution of permuted statistics might not be a random sample of the full permutation distribution. It remains to be seen whether the restricted distribution differs enough to bias the estimated *p* values.
- Whenever a random shuffle results in an unequal number of observed categories across groups, the categories could be collapsed before fitting the model to that permutation. This is not recommended because the model *df* would vary across permutations, so the expected value (even if biased due to violated assumptions) would not match that of the original data.
- A researcher could proactively collapse sparse categories before fitting their model to the original data. This could advantageously avoid the problem of restricted resampling or mismatched *df* in methods above; in fact, it might be the only reasonable option if the original data already had an unequal number of observed categories across groups (assuming the software package could not allow different numbers of thresholds across groups). The disadvantage, though, would be ignoring some information about individual differences.

We employed the second method (reshuffling) in our current investigation because it is the simplest to implement and because the reason for reshuffling to make a new permutation (i.e., sparseness in a permuted data set) is itself a random process that occurs only when the grouping variable is shuffled to begin with. We therefore expected that the restricted permutation distribution under the second method above should provide *p* values that yield nominal Type I error rates. A more thorough investigation comparing the methods described above is beyond the scope of the current article, but is suggested for future research.

APPENDIX B

The lavTestLRT function in R with the argument A.method = "exact" initially showed near-zero Type I error rates and very low power (see Table B1). These results were gathered before adding the unbalanced sample size conditions.

TABLE B1
Scalar Invariance Type I Errors and Power

Invariance	*Sample Size*	*Responses*	*Threshold Symmetry*	*Rejection Rate*
Yes	150	2	Symmetric	.004
Yes	300	2	Symmetric	.002
Yes	150	5	Symmetric	.000
Yes	300	5	Symmetric	.000
Yes	150	2	Non-Symmetric	.010
Yes	300	2	Non-Symmetric	.004
Yes	150	5	Non-Symmetric	.001
Yes	300	5	Non-Symmetric	.000
No	150	2	Symmetric	.084
No	300	2	Symmetric	.220
No	150	5	Symmetric	.091
No	300	5	Symmetric	.347
No	150	2	Non-Symmetric	.070
No	300	2	Non-Symmetric	.143
No	150	5	Non-Symmetric	.049
No	300	5	Non-Symmetric	.226

Note. Rejection rates are provided for α= .05.