Article

Educational and Psychological Measurement
2018, Vol. 78(2) 272–296

DOI: 10.1177/0013164416683754
journals.sagepub.com/home/epm

The Impact of Model Parameterization and Estimation Methods on Tests of Measurement Invariance With Ordered Polytomous Data

Natalie A. Koziol[1] and James A. Bovaird[1]

Abstract

Evaluations of measurement invariance provide essential construct validity evidence—a prerequisite for seeking meaning in psychological and educational research and ensuring fair testing procedures in high-stakes settings. However, the quality of such evidence is partly dependent on the validity of the resulting statistical conclusions. Type I or Type II errors can render measurement invariance conclusions meaningless. The present study used Monte Carlo simulation methods to compare the effects of multiple model parameterizations (linear factor model, Tobit factor model, and categorical factor model) and estimators (maximum likelihood [ML], robust maximum likelihood [MLR], and weighted least squares mean and variance-adjusted [WLSMV]) on the performance of the chi-square test for the exact-fit hypothesis and chi-square and likelihood ratio difference tests for the equal-fit hypothesis for evaluating measurement invariance with ordered polytomous data. The test statistics were examined under multiple generation conditions that varied according to the degree of metric noninvariance, the size of the sample, the magnitude of the factor loadings, and the distribution of the observed item responses. The categorical factor model with WLSMV estimation performed best for evaluating overall model fit, and the categorical factor model with ML and MLR estimation performed best for evaluating change in fit. Results from this study should be used to inform the modeling decisions of applied researchers. However, no single analysis

[1]University of Nebraska–Lincoln, Lincoln, NE, USA

Corresponding Author:
Natalie A. Koziol, Nebraska Center for Research on Children, Youth, Families and Schools, University of Nebraska–Lincoln, 303 Mabel Lee Hall, Lincoln, NE 68588-0235, USA.
Email: nkoziol@unl.edu

combination can be recommended for all situations. Therefore, it is essential that researchers consider the context and purpose of their study.

Keywords

categorical confirmatory factor analysis, differential item functioning, limited information estimation, measurement invariance, polytomous data, robust maximum likelihood, Tobit model

Measurement invariance (MI) is indicated when a latent construct has equivalent measurement properties across multiple groups or time points. That is, conditioning on the latent construct, MI is indicated if the distribution of observed responses is equivalent across groups/time points (Kim & Yoon, 2011). Evaluations of MI provide essential construct validity evidence—a prerequisite for seeking meaning in psychological and educational research and ensuring fair testing procedures in high-stakes settings (Brown, 2006). For example, multiple-group analyses such as cross-cultural comparisons are meaningful only to the extent that the same construct is being measured across groups. Likewise, longitudinal evaluations depend on the assumption that the same construct is being measured across time. The use of test scores for making high-stakes decisions, such as employment or credentialing decisions, raises the question of fairness. Evaluations of MI are performed to flag items that function differently across groups so the items can be further examined for possible bias or unfairness.

From a factor-analytic perspective, evaluations of MI involve four primary steps (Brown, 2006). Step 1 is to evaluate configural invariance, which is demonstrated when the same factor structure holds across groups (i.e., the same dimensionality holds, and the same items load on each factor). Assuming configural invariance, Step 2 is to evaluate metric invariance, a test of whether the proportion of true score variance is equivalent across groups (i.e., whether the item factor loadings are equivalent across groups). In the item response theory (IRT) literature, metric noninvariance is referred to as nonuniform differential item functioning (DIF), which indicates an interaction between the latent construct and group membership (Narayanan, 1996). Contingent on evidence of metric invariance, Step 3 is to evaluate scalar invariance to determine whether the item intercepts are equivalent across groups. Noninvariance at this step indicates a main effect of group membership on item responses, referred to as uniform DIF in IRT (Narayanan, 1996). Finally, Step 4 is to evaluate residual variance invariance to determine whether the proportion of error variance is equivalent across groups. The present study focuses on the first two steps of MI, as historically these steps have garnered the most attention (Millsap & Olivera-Aguilar, 2012). Under the commonly used free baseline approach (Wang, Tay, & Drasgow, 2013), configural invariance is the foundation for all other tests of MI, where evaluations of MI cannot proceed without it. In turn, metric invariance is often considered to be the

most important indicator of MI (Meade & Bauer, 2007; Suh & Cho, 2014), as equal true score variance suggests that the same construct is being measured in both groups (Meredith & Horn, 2001).

There are many statistical methods and frameworks for evaluating MI, which have been extensively studied and compared (e.g., Narayanan, 1996; Wang et al., 2013; Woods, 2011). For this study, we consider the chi-square test (T) for the exact-fit hypothesis for testing configural invariance, and the chi-square and likelihood ratio difference statistics (ΔT and ΔG^2, respectively) for the equal-fit hypothesis for testing metric invariance, within the confirmatory factor analysis (CFA) framework. Configural invariance is demonstrated if an unconstrained multiple-group model demonstrates acceptable fit as indicated by a failure to reject the exact-fit hypothesis (Brown, 2006). Using a free baseline approach, metric invariance is demonstrated if model fit does not get significantly worse when constraining the factor loadings to be equal across groups as indicated by a failure to reject the equal-fit hypothesis (Brown, 2006).

Although approximate fit indexes have been studied in the context of measurement invariance (e.g., Cheung & Rensvold, 2002; Fan & Sivo, 2009; French & Finch, 2006; Meade & Bauer, 2007; Meade, Johnson, & Braddy, 2008; Sass, Schmitt, & Marsh, 2014), we focus solely on test statistics (T, ΔT, and ΔG^2) due to serious limitations with using approximate fit indexes to make binary decisions about model fit. In particular, research has shown that suggested thresholds for ''acceptable'' model fit do not generalize across modeling and sampling contexts (Kline, 2016). Barrett (2007) goes as far as saying, ''I would now recommend banning *ALL* such indices from ever appearing in any paper as indicative of model 'acceptability' or 'degree of misfit''' (p. 821). While test statistics have been criticized as being overly sensitive to small deviations in fit when the sample size is large, Hayduk, Cummings, Boadu, Pazderka-Robinson, and Boulianne (2007) refute this criticism and go on to argue that ''χ^2, degrees of freedom, and the associated probability must be reported for all manuscripts reporting SEM results'' (p. 845).

The above framework centers on global assessments of invariance—assessments that ''measure only average or overall model-data correspondence'' (Kline, 2016, p. 262). Whereas global assessments can be used to identify initial support for a model, researchers should conduct follow-up analyses to identify areas of local misfit (i.e., misspecification of specific pathways) and potential sources of noninvariance (Kline, 2016). Both global and localized information add to the larger body of psychometric evidence, and thus both should be reported. We focus on global assessments because they are so widely used. To substantiate this claim, we searched the PsycARTICLES database for articles related to invariance (using the search phrase: invariance) that were published in 2015 or 2016 or were available through advance online publication. Of the 79 articles that described an application of MI, only 2 did not present information about global model fit or change in model fit. This brief review indicates that global fit information continues to be widely referenced in practice and thereby warrants additional methodological research.

Although MI is a widely studied topic, comparatively little research has investigated the impact of model parameterization and estimation decisions on such evaluations, particularly in the context of modeling ordered polytomous data. To partially fill this void, the present study used Monte Carlo simulation methods to compare the effects of multiple model parameterizations (linear factor model [LFM], Tobit factor model [TFM], and categorical factor model [CFM]) and estimators (maximum likelihood [ML], robust maximum likelihood [MLR], and weighted least squares mean and variance-adjusted [WLSMV]) on the performance of T, ΔT, and ΔG^2 for evaluating MI with ordered polytomous data. The test statistics were examined under multiple generation conditions that varied according to the degree of metric noninvariance, the size of the sample, the magnitude of the factor loadings, and the distribution of the observed item responses. Before discussing our methods and results, we first detail the model parameterizations and estimators of interest and then review the relevant literature.

Measurement Models

Modern measurement frameworks such as CFA and IRT attempt to explain patterns of observed item responses by statistically regressing the observed responses on latent person and item characteristics (de Ayala, 2009). Within these broad frameworks, many statistical models can be conceptualized depending on the nature of the data. As such, it is not always clear which model is most appropriate for a given context. Of particular interest to the present study is the analysis of ordered polytomous response data in which items are rated on a 5-point scale, the most commonly used scale for Likert-type items (Lozano, García-Cueto, & Muñiz, 2008). Although such data are clearly discrete and often bound by floor or ceiling effects, it is common practice to treat the data as if they are continuous and normally distributed (Lubke & Muthén, 2004), despite the fact that alternative approaches may be more theoretically appropriate. Three possible modeling frameworks are described below, including the commonly employed LFM that assumes normality of responses. For simplicity of presentation, a single latent factor is assumed throughout.

Linear Factor Model

As the name suggests, the LFM posits a linear relationship between the observed item responses and latent person and item characteristics:

$$X_i = \nu_i + \lambda_i \xi + \delta_i, \tag{1}$$

where X_i is the response to item i by a randomly chosen person from the population, ν_i is the intercept for item i (the expected response to item i of a randomly chosen person with an average level of the latent factor), λ_i is the factor loading for item i (the expected change in the response to item i associated with a one unit increase in the latent factor score), ξ is the latent (person) factor, and δ_i is the residual for item i

where $\delta_i \sim N(0, \theta_i)$ (Bollen, 1989). For $\xi \sim N(0, 1)$, it is assumed that $X_i \sim N\left(\nu_i, \lambda_i^2 + \theta_i\right)$.

Tobit Factor Model

The LFM assumes data are continuous and normally distributed. In psychological and educational research, floor and ceiling effects occur, which prevent part of the distribution of responses from being observed. For example, many cognitive instruments are designed to differentiate among average ability students. Administering the instruments to a sample of gifted students will result in an excess of responses at the upper end of the scale (McBee, 2010). This is assumed to be a limitation of the instrument and not an indication of the true distribution of cognitive abilities—if a more appropriate instrument was administered, it is assumed that the responses would follow a normal distribution.

When there is an excess of responses at one or both ends of the scale, it suggests that the true responses are censored (i.e., only a lower or upper bound can be observed) for persons with particularly low or high levels of the latent construct. In this case, a more theoretically correct alternative to the LFM is the TFM (Tobin, 1958), which assumes that the observed responses are continuous and follow a censored normal distribution. A general representation of the model that allows for censoring at both ends of the scale is given by

$$X_i = \begin{cases} \tau_L & \text{if} \quad X_i^* \leq \tau_L \\ X_i^* = \nu_i + \lambda_i \xi + \delta_i & \text{if} \quad \tau_L < X_i^* < \tau_U \\ \tau_U & \text{if} \quad X_i^* \geq \tau_U \end{cases}, \tag{2}$$

where X_i is the observed response to item i, X_i^* is the true (uncensored) response to item i, and τ_L and τ_U are the lower and upper bounds of the observed data, respectively. For $\xi \sim N(0, 1)$, it is assumed that $X_i^* \sim N\left(\nu_i, \lambda_i^2 + \theta_i\right)$. Equation (2) shows that unlike the LFM, the TFM makes a distinction between the true response and the observed response.

Categorical Factor Model

The LFM and TFM assume that the observed responses are continuous and follow a normal or censored normal distribution, respectively. Neither of these models accounts for the discrete nature of Likert-type data. A third option for analyzing ordered polytomous data is to specify a model that explicitly accounts for this discreteness. Under the IRT framework, one such model is the graded response model (GRM; Samejima, 1969).[1] The present study focuses on a parameterization of the GRM that specifies factor loadings and thresholds (see, e.g., IRT in M*plus*, 2013, p. 1), given as

$$P(X_i = c|\xi) = \int_{\tau_{ic-1}}^{\tau_{ic}} \psi(z|\xi)\mathrm{d}z = \frac{\exp(-\lambda_i\xi + \tau_{ic})}{1 + \exp(-\lambda_i\xi + \tau_{ic})} - \frac{\exp(-\lambda_i\xi + \tau_{ic-1})}{1 + \exp(-\lambda_i\xi + \tau_{ic-1})}, \quad (3)$$

where c is the response option (e.g., $c = 1, \ldots, 5$ for a 5-option scale), τ_i is the set of $C - 1$ thresholds for item i, ψ is the logistic probability density function (pdf), and all other terms are defined above. If c is the first response option then the second term in Equation (3) goes to 0, and if c is the last response option then the first term is equal to 1. Alternatively, Equation (3) can be specified in terms of the standard normal pdf (φ) and corresponding cumulative distribution function (Φ):

$$P(X_i = c|\xi) = \int_{\tau_{ic-1}}^{\tau_{ic}} \varphi(z|\xi)\mathrm{d}z = \Phi(-\lambda_i\xi + \tau_{ic}) - \Phi(-\lambda_i\xi + \tau_{ic-1}). \quad (4)$$

Estimators

Another modeling consideration is the model estimator. A major distinction among estimators is their classification as full-information or limited-information methods. Full-information methods are more efficient than limited-information methods, but the latter are computationally faster and may be the only feasible option when estimating complex models (Forero & Maydeu-Olivares, 2009).

The present study focuses on ML estimation, a full-information method, and WLSMV estimation, a limited-information method. In addition, a robust ML estimator (referred to as MLR in M*plus*; L. K. Muthén & Muthén, 1998-2015) is examined that offers Huber–White sandwich variance estimators and corrected test statistics for use under nonideal sample conditions. In the context of performing MI analyses within M*plus*, ML and MLR can be used to estimate all three measurement models described in the previous section, whereas WLSMV can only be used to estimate the CFM.

Maximum Likelihood

For the LFM, ML estimation is performed by minimizing a fit function derived from a multivariate normal distribution (see Bollen, 1989):

$$F_{ML}(\boldsymbol{\theta}) = \mathbf{log}|\boldsymbol{\Sigma}(\boldsymbol{\theta})| + \mathbf{tr}\left[\boldsymbol{S}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\right] - \mathbf{log}|\boldsymbol{S}| - q, \quad (5)$$

where $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the model-implied covariance matrix, $\boldsymbol{S}$ is the sample covariance matrix, and q is the number of items. A test of the exact-fit hypothesis is obtained by comparing

$$T_{ML} = (N - 1)F_{ML}\left(\hat{\boldsymbol{\theta}}\right), \quad (6)$$

to a critical value from a chi-square distribution with $q/(2[q + 1]) - t$ degrees of freedom, where t is the number of estimated parameters. To evaluate difference in fit

between a constrained model (M_0) and unconstrained model (M_1), a chi-square difference test of the equal-fit hypothesis can be performed by comparing

$$\Delta T_{ML} = (N-1)F_{ML}\left(\hat{\boldsymbol{\theta}}_0\right) - (N-1)F_{ML}\left(\hat{\boldsymbol{\theta}}_1\right), \tag{7}$$

to a critical value from a chi-square distribution with $t_1 - t_0$ degrees of freedom.

The sample covariance matrix contains only information about the items' means and covariances; when the assumption of normality is not tenable, as in the TFM and CFM cases, the sample covariance matrix does not sufficiently summarize the data. Instead of minimizing a fit function, ML estimation is carried out by maximizing the corresponding likelihood function given the observed sample response patterns (Forero & Maydeu-Olivares, 2009). For the CFM, it is possible to calculate Pearson's χ^2 statistic using the frequencies from the joint contingency table; yet the sparseness that occurs when there are a large number of items and item categories violates the underlying multinomial distribution assumption (Maydeu-Olivares & Cai, 2006). As such, the χ^2 statistic should not be used to assess global fit under these conditions. However, for both the TFM and CFM, change in model fit can be evaluated by comparing

$$\Delta G^2 = -2(\log(L_0) - \log(L_1)), \tag{8}$$

to a chi-square distribution with $t_1 - t_0$ degrees of freedom.

A further complication with ML in the case of the TFM and CFM is that there is no closed-form solution for integrating over the latent (person) variable distribution. Integration is the basis for marginal ML estimation (MMLE; note that we will simply refer to MMLE as ML). For these models, a numerical integration algorithm is required to approximate the marginal distribution (de Ayala, 2009).

Robust Maximum Likelihood

MLR uses the same maximization procedure as ML but offers empirically adjusted variance estimators and test statistics (L. K. Muthén & Muthén, 1998-2015). These statistics are generally preferred to their model-based counterparts when model-based assumptions (e.g., normality, large sample sizes) are violated (Satorra & Bentler, 2001). For the LFM, the corrected test statistic is calculated as

$$T_{MLR} = T_{ML}/c, \tag{9}$$

where Yuan and Bentler (2000) provide details on the calculation of the correction factor (c). M*plus* uses a slightly different calculation for c, but L. K. Muthén and Muthén (1998-2015) note that $T_{MLR-LFM}$ is ''asymptotically equivalent to the Yuan-Bentler T2* test statistic'' (p. 603). The corresponding adjusted chi-square and likelihood ratio difference tests are calculated as (see Satorra & Bentler, 2001)

$$\Delta T_{MLR} = (T_{ML,0}c_0 - T_{ML,1}c_1)/c_d, \tag{10}$$

and

$$\Delta G^2_{MLR} = -2\left(G^2_0 - G^2_1\right)/c_d, \tag{11}$$

respectively, where $c_d = (t_0 c_0 - t_1 c_1)/(t_0 - t_1)$.

Weighted Least Squares Mean and Variance-Adjusted

WLSMV is a limited-information estimator because it uses only first- and second-order moment information. Like ML for the LFM, WLSMV relies on a least squares approach to minimize a fit function given as

$$F_{WLSMV}(\boldsymbol{\theta}) = (\boldsymbol{s} - \boldsymbol{\sigma}(\boldsymbol{\theta}))^T \boldsymbol{W}^{-1} (\boldsymbol{s} - \boldsymbol{\sigma}(\boldsymbol{\theta})), \tag{12}$$

where $\boldsymbol{s}$ contains the sample thresholds and polychoric correlations, $\boldsymbol{\sigma}(\boldsymbol{\theta})$ contains the model-implied thresholds and polychoric correlations, and $\boldsymbol{W}$ is a diagonal weight matrix (B. O. Muthén, du Toit, & Spisic, 1997). Polychoric correlations are used here because it is assumed that latent, normally distributed responses underlie the observed ordinal item responses. A mean- and variance-adjusted test of the exact-fit hypothesis is then calculated as

$$T_{WLSMV} = N \times F_{WLSMV}\left(\hat{\boldsymbol{\theta}}\right) \times \left(d/\mathbf{tr}(\boldsymbol{U}\Gamma)^2\right), \tag{13}$$

where the full calculations for d, $\boldsymbol{U}$, and Γ are provided by B. O. Muthén et al. (1997). The corresponding difference test for the equal-fit hypothesis is calculated as

$$\Delta T_{WLSMV} = (T_{WLSMV,0} - T_{WLSMV,1}) \times (d/tr(M)), \tag{14}$$

where the full calculations for d and M are provided by Asparouhov and Muthén (2006).

Past Research

A handful of studies have examined the performance of the aforementioned test statistics in the context of modeling approximately continuous ordered polytomous data (i.e., data in which the items are based on a 5-point scale). We focus on research involving 5-point scales, not only because 5-point scales are the most commonly used (Lozano et al., 2008), but because there seems to be little empirical or theoretical support for treating data based on fewer response categories as continuous (e.g., Sass et al., 2014).

Overall Model Fit

With respect to the test of the exact-fit hypothesis, the focus has been on T_{ML-LFM} and $T_{WLSMV-CFM}$. For a single-factor model across varying sample sizes, Babakus,

Ferguson, and Jöreskog (1987) and B. Muthén and Kaplan (1985) found that the Type I error rate of T_{ML-LFM} became increasingly inflated as the distribution of observed responses became increasingly skewed. The biggest limitation of B. Muthén's and Kaplan's study is that the results were based on only 25 replications, making it difficult to disentangle variability due to the study factors from simulation error. Although Babakus et al. relied on a greater number of replications, they collapsed their results across factor loading conditions, thereby overlooking a potentially important study factor. Importantly, neither study considered power. Power is a particularly important consideration for high-stakes MI applications in which noninvariance presents a source of bias. Finally, both studies were conducted in the context of a single-group analysis, and thus, it remains unclear whether the results generalize to the context of testing MI.

Lubke and Muthén (2004) considered a multiple-group analysis of a single-factor model and observed that the T_{ML-LFM} Type I error rate was particularly inflated when the sample size and factor loadings were large, and the items were skewed. Unfortunately, the item distribution condition (i.e., all items bell-shaped vs. items skewed in varying directions) was confounded in that it is unclear whether inflation was due to the skewness itself, or due to the fact that the distribution varied across items. In addition, the interpretability of results was limited by relatively large amounts of simulation error (only 100 replications were used) and lack of attention to power.

Three studies have considered the performance of $T_{WLSMV-CFM}$ in the context of modeling approximately continuous ordered polytomous data. Flora and Curran (2004) found that for a single-group analysis, $T_{WLSMV-CFM}$ was too liberal for the smallest sample size (N = 100), particularly when the model was complex. The authors did not consider power, however, and called for future research to address this limitation. The study by Kim and Yoon (2011) partially filled this gap. Their results, based on a multiple-group single-factor model, indicated that the $T_{WLSMV-CFM}$ Type I error rate was acceptable across sample size conditions (n = 200 or 500 per group). Power to detect model misfit exceeded .80 only when the degree of misfit was large. Although these results provide some insight into the power of $T_{WLSMV-CFM}$, the relative nature of power and the study's lack of comparison test statistic make it difficult to meaningfully interpret the results. Notably, Beauducel and Herzberg (2006) directly compared the performance of T_{ML-LFM} and $T_{WLSMV-CFM}$. For a single-group analysis of data based on a 5-point scale, results indicated that both test statistics had inflated Type I error rates, but T_{ML-LFM} was slightly more liberal when the factors of the multidimensional model were uncorrelated and $T_{WLSMV-CFM}$ was slightly more liberal when the factors were correlated. Unfortunately, results were averaged across sample sizes and degrees of model complexity, preventing more nuanced conclusions. Lack of attention to power also limited the comparison.

Past research provides some insight into the effects of model parameterization and estimation decisions on the performance of the test for the exact fit hypothesis, but there remain important gaps in the literature. In particular, the performance of

$T_{MLR-LFM}$ when analyzing ordered polytomous data needs to be evaluated. Likewise, direct comparisons among T_{ML-LFM},$T_{MLR-LFM}$, and $T_{WLSMV-CFM}$ are needed. Power should be evaluated in addition to the Type I error rate, as power is especially critical when assessing MI. Relatedly, more research is needed to determine whether previous findings generalize to the multiple-group MI context. Finally, the impact of factor loading magnitude and distribution of the observed ordinal variables on the performance of $T_{WLSMV-CFM}$ needs to be investigated.

Change in Model Fit

Two studies have performed comparisons involving the chi-square and/or likelihood ratio difference test for the equal-fit hypothesis that are particularly relevant to the present study. Kim and Yoon (2011) compared ΔG^2_{ML-CFM} and $\Delta T_{WLSMV-CFM}$ in the context of simultaneously evaluating metric and scalar invariance using a Bonferroni-corrected constrained-baseline approach where each item was tested individually. Their results indicated that the Type I error rate (i.e., the number of times model fit significantly improved upon freeing a non-DIF item) for $\Delta T_{WLSMV-CFM}$ was generally inflated, particularly for larger sample sizes and increased DIF. A similar pattern held for ΔG^2_{ML-CFM}, but the magnitude of inflation was smaller. Holding the Type I error rates approximately equal, $\Delta T_{WLSMV-CFM}$ generally had greater power than ΔG^2_{ML-CFM}. The biggest limitation of this study is the use of a constrained-baseline approach for detecting DIF, as this approach has been found to produce considerably inflated Type I error rates when compared to a free-baseline approach (Stark, Chernyshenko, & Drasgow, 2006).

Using a free-baseline approach, Sass et al. (2014) compared the performance of ΔT_{ML-LFM}, $\Delta T_{MLR-LFM}$, and $\Delta T_{WLSMV-CFM}$ for simultaneously evaluating metric and scalar invariance within the context of a multiple-group single-factor model. They found that ΔT_{ML-LFM} generally held the correct size (i.e., the Monte Carlo estimated Type I error rate was approximately equal to $\alpha)^2$ across sample sizes, regardless of whether the observed responses followed a symmetric or asymmetric distribution, whereas $\Delta T_{MLR-LFM}$ was overly conservative when the distribution was asymmetric. For small sample sizes, $\Delta T_{WLSMV-CFM}$ was overly liberal regardless of the distribution of responses. Holding the Type I error rates approximately constant, $\Delta T_{WLSMV-CFM}$ had the greatest power to detect metric noninvariance. As expected based on the Type I error results, $\Delta T_{MLR-LFM}$ was considerably less powerful than the other test statistics under the asymmetric condition.

The findings of Kim and Yoon (2011) and Sass et al. (2014) provide initial insight into the effects of model parameterization and estimation decisions on the performance of the chi-square and likelihood ratio difference tests when analyzing approximately continuous ordered polytomous data. However, additional research is needed in this area. Evaluations of the Tobit model using ΔG^2_{ML-TFM} and $\Delta G^2_{MLR-TFM}$ are nonexistent. Given the regularity of floor and ceiling effects when studying special populations (e.g., gifted students), the Tobit model should be investigated as an

alternative to the LFM. Comparisons of the LFM and CFM using full-information estimators are also needed. In addition, comparisons of limited- and full-information estimators for the CFM should be conducted using a free-baseline approach. Finally, the effect of factor loading magnitude on the performance of the test statistics needs to be evaluated. The present study seeks to fill these gaps.

Method

Data Generation

Our aforementioned review of recently published articles describing applications of MI indicated considerable variability in the types of factor models being evaluated. For example, across articles the number of factors ranged from 1 to 11, and the number of items per factor ranged from 2 to 27. Likewise, item parameter estimates varied widely. It was therefore impossible to design a study that would generalize across all scenarios. However, to strengthen the generalizability of our results, we attempted to model our simulation after that of previous research and based on general guidelines offered in the literature on latent variable modeling. As we note below, in a few instances we sacrificed some external validity in favor of minimizing confounding variability, which would have threatened the internal validity of our results.

In the first phase of data generation, the Monte Carlo procedure in M*plus* Version 5.2 was used to generate multiple-group multivariate normal data from a 10-item (cf., Flora & Curran, 2004; Sass et al., 2014) single-factor model (cf., Babakus et al., 1987; Kim & Yoon, 2011; Lubke & Muthén, 2004; B. Muthén & Kaplan, 1985; Sass et al., 2014) specified as

$$X_{ig} = \nu_i + \lambda_{ig}\xi + \delta_{ig}, \tag{15}$$

where $g = 1, 2$ denotes group membership, $\xi \sim N(0, 1)$ for both groups, and all other terms are defined above. Item factor loadings were manipulated as a condition in the study, but across conditions, the total variance for each item was fixed at one by specifying the item residual variances to be $\theta_{ig} = 1 - \lambda_{ig}^2$. Item intercepts were specified to be zero for all items within and across groups. In the second phase of data generation, performed in SAS Version 9.2, the continuous normally distributed data were converted into ordered 5-category items by imposing a set of categorization thresholds (cf., Babakus et al., 1987; Flora & Curran, 2004; Lubke & Muthén, 2004; B. Muthén & Kaplan, 1985):

$$\tilde{X}_{ig} = \begin{cases} 0 & \text{if} \quad X_{ig} < \tau_1 \\ 1 & \text{if} \quad \tau_1 \leq X_{ig} < \tau_2 \\ 2 & \text{if} \quad \tau_2 \leq X_{ig} < \tau_3 \\ 3 & \text{if} \quad \tau_3 \leq X_{ig} < \tau_4 \\ 4 & \text{if} \quad X_{ig} \geq \tau_4 \end{cases}, \tag{16}$$

where the thresholds were manipulated as a condition in the study.

Four sample design factors were manipulated, including the degree of metric noninvariance, size of the sample, magnitude of factor loadings, and distribution of the observed responses. Each of these factors is described below.

Three levels of metric noninvariance were considered: (1) 10 invariant items and 0 noninvariant items (invariance condition); (2) 8 invariant items and 2 items exhibiting moderate noninvariance ($z_{\lambda_{ig}} - z_{\lambda_{ig'}} = .32$, where z is Fisher's z-transformation of the factor loading); and (3) 8 invariant items and 2 items exhibiting large noninvariance ($z_{\lambda_{ig}} - z_{\lambda_{ig'}} = .60$). Specifying 20% of the items to be noninvariant is in line with past MI research (cf., French & Finch, 2006; Sass et al., 2014).

The total sample size was either 400 or 1,000, evenly balanced across groups ($n_1 = n_2 = 200$ or 500). We chose group sizes of 200 and 500 because they have been suggested as rough guidelines for the minimum sample size necessary to perform latent variable modeling (e.g., Barrett, 2007; de Ayala, 2009).

The magnitude of factor loadings was also manipulated. For the invariant condition, $\lambda_{ig} = .5$, .7, or .9. These represent standardized loadings, as the total variance for each item was fixed at 1. Kline (2016) recommends using items with standardized loadings of at least .7, so the $\lambda_{ig} = .7$ condition corresponds to a minimally adequate setting, whereas the $\lambda_{ig} = .5$ and $\lambda_{ig} = .9$ conditions correspond to inadequate and more than adequate settings, respectively. We note as a limitation that the assumption of tau equivalence may not be realistic in many practical applications. However, we felt this restriction was necessary to minimize confounding differences across conditions due to nonmanipulated differences in true score variability. For the noninvariant conditions, $\lambda_{ig} = .7$ for all 10 items in the reference group and for the 8 invariant items in the focal group, and $\lambda_{ig} = .5$ or .9 for the 2 items in the focal group exhibiting moderate and large noninvariance, respectively.

Finally, we manipulated the distribution of observed responses by imposing different sets of categorization thresholds: (1) Approximately normal distribution based on the set of thresholds, −1.645, −0.643, 0.643, 1.645, resulting in Response Options 0 to 4 being endorsed on average by 5%, 21%, 48%, 21%, and 5% of the sample, respectively (based on B. Muthén & Kaplan, 1985), and resulting in skew = .00 and kurtosis = .02; (2) Approximately normal distribution but censored from below based on the set of thresholds, −0.842, 0.050, 0.995, 1.960, resulting in Response Options 0 to 4 being endord on average by 20%, 32%, 32%, 13.5%, and 2.5% of the sample, respectively, and resulting in skew = .25 and kurtosis = −.59; and (3) L-shaped distribution based on the set of thresholds, 0.000, 0.527, 0.845, 1.290, resulting in Response Options 0 to 4 being endorsed on average by 50%, 20%, 10%, 10%, and 10% of the sample, respectively, and resulting in skew = .98 and kurtosis = −.38. The approximately normal and censored distributions were chosen to reflect situations in which the assumption of a normal or censored normal distribution, respectively, may be practical despite the discreteness of the data. The L-shaped distribution was chosen to reflect a situation in which the data are skewed but the assumption of a normal or censored normal distribution is implausible. We acknowledge as a limitation that these thresholds are arbitrary in the sense that they

are not estimated from actual data. Ultimately, we felt it was most informative to evaluate the LFM and TFM under ideal situations (our reasoning being that if they do not perform well under ideal situations then it is unlikely they will perform well under nonideal situations), which necessitated purposeful selection of thresholds.

Crossing the design factors resulted in a total of 30 data generation conditions. Ten thousand samples, or as many samples as necessary to obtain 10,000 samples with all 5 response categories being endorsed at least once by both groups for all 10 items, were generated for each condition. This constraint was applied to ensure comparability across estimators, as the use of WLSMV with the grouping option in M*plus* requires that both groups endorse the same response options.

Data Analysis

Data were analyzed using multiple combinations of models and estimators. The three measurement models included the LFM, TFM, and CFM. The three estimators included ML, MLR, and WLSMV. Default link functions were used in conjunction with the CFM—the logit link for ML and MLR and the probit link for WLSMV. ML and MLR were fully crossed with each of the measurement models, but WLSMV was paired only with the CFM. This resulted in seven different combinations of models and estimators.

Combining the generation and analysis conditions produced a total of 210 study cells. All 210 cells were evaluated in M*plus* under each of two MI hypotheses: configural and metric invariance. In testing configural invariance, item parameters were freely estimated across groups. The latent factor was identified in each group by fixing the mean at zero and the variance at one. In testing metric invariance, factor loadings were constrained to be equal across groups resulting in 9 additional degrees of freedom (10 constrained factor loadings minus the corresponding freely estimated factor variance for Group 2).

Outcome Measures

Outcome measures included the Monte Carlo estimated Type I error rate and power of the chi-square tests of the exact-fit hypothesis (T_{ML-LFM}, $T_{MLR-LFM}$, and $T_{WLSMV-CFM}$) and chi-square and likelihood ratio difference tests of the equal-fit hypothesis (ΔT_{ML-LFM}, $\Delta T_{MLR-LFM}$, ΔG^2_{ML-TFM}, $\Delta G^2_{MLR-TFM}$, ΔG^2_{ML-CFM}, $\Delta G^2_{MLR-CFM}$, and $\Delta T_{WLSMV-CFM}$). The nominal Type I error rate for both sets of test statistics was .05. The tests of the exact-fit hypothesis were based on the constrained (metric invariance) model. While overall fit is most relevant to testing the hypothesis of configural invariance, assessing overall fit of the metric invariance model allowed us to evaluate both the Type I error rate and power. The difference tests were calculated using a free-baseline approach in which the fully constrained metric invariance model was compared to the unconstrained configural model.

Results

Summary of Replications

Due to the simplicity of the factor model, lack of model misspecification within groups, and adequacy of the overall sample size, nonconvergence and inadmissible solutions were a nonissue. However, as previously noted, samples were discarded if not all five response categories were endorsed at least once for each item in each group. Table 1 shows the total number of replications required to obtain 10,000 analysis replications. As expected, the conditions in which the sample size was small and the responses followed a censored distribution required the greatest number of replications to be generated. This is due to the fact that the average response frequency was set to 2.5% for the final response option under this condition and was therefore more likely to be unobserved for any given replication. Very few additional replications were required under all other conditions. The general implication of discarding replications is that the data for the analysis replications were, on average, *slightly* less sparse than the data for the complete set of generation replications. As such, the impact of the categorization thresholds on the performance of the test statistics may have been slightly attenuated.

Overall Model Fit

See Table 2 and Figure 1 for the Monte Carlo estimated Type I error rate and power of the test of the exact-fit hypothesis across generation and analysis conditions. For the invariant condition, using a normal approximation to the binomial,

$$.05 \pm 1.96 \times \sqrt{\frac{.05 \times (1 - .05)}{10,000}}, \tag{17}$$

it is expected (with 95% confidence interval) that test statistics with a true size of .05 will have an estimated size between .0457 and .0543. Following the approach of Sass et al. (2014), we use a more liberal criterion of two standard errors (between .0415 and .0585) to evaluate the acceptability of the test statistics. Type I error rates outside this range are boldfaced in the table.

Linear Factor Model. T_{ML-LFM} was too liberal across all generation conditions. The Type I error rate was most noticeably inflated for the conditions in which the responses followed an L-shaped distribution, with inflation reaching as high as .931 under the small sample size, large factor loading condition. The error rate was much less pronounced for the approximately normal and censored conditions, where these conditions had very similar effects on the error rate. Across all three response distribution conditions, inflation increased with decreased sample size and increased magnitude of the factor loadings. $T_{MLR-LFM}$ was also consistently liberal, but compared to ML, the Type I error rate was considerably less inflated for the L-shaped

Table 1. Total Number of Replications Generated to Obtain 10,000 Analysis Replications.

Outcome	n_1/n_2	τ	λ	Total replications
Type I error rate (invariance)	200/200	N	.5	10,012
			.7	10,013
			.9	10,012
		C	.5	11,332
			.7	11,295
			.9	11,121
		L	.5	10,000
			.7	10,000
			.9	10,000
	500/500	N	.5	10,000
			.7	10,000
			.9	10,000
		C	.5	10,002
			.7	10,001
			.9	10,000
		L	.5	10,000
			.7	10,000
			.9	10,000
Power (noninvariance)	200/200	N	.5	10,017
			.9	10,019
		C	.5	11,304
			.9	11,339
		L	.5	10,000
			.9	10,000
	500/500	N	.5	10,000
			.9	10,000
		C	.5	10,002
			.9	10,002
		L	.5	10,000
			.9	10,000

Note. n_1/n_2 = sample size of Group 1/Group 2. τ = distribution of responses (N = approximately normal, C = censored, L = L-shaped). λ = factor loading for all 10 invariant items under the full metric invariance condition or for the 2 noninvariant items in the focal group under the metric noninvariance condition (where the loadings for the other items were .7).

distribution condition. Across factor loading and response distribution conditions, inflation decreased with increased sample size. Inflation increased with increased magnitude of the factor loadings but only for the L-shaped distribution condition.

Categorical Factor Model. $T_{WLSMV-CFM}$ held the correct size when the magnitude of the factor loadings was large, but the error rate was inflated when the factor loadings were of a smaller magnitude. The sample size and response distribution conditions had a noticeable (albeit small) effect on the error rate for the smaller factor loading

Table 2. Monte Carlo Estimated Type I Error Rate and Power of the Chi-Square Test of the Exact-Fit Hypothesis for Evaluating Metric Invariance.

				LFM		CFM
Outcome	n_1/n_2	τ	λ	ML	MLR	WLSMV
Type I error rate (invariance)	200/200	N	.5	**0.070**	**0.085**	**0.107**
			.7	**0.073**	**0.078**	**0.079**
			.9	**0.127**	**0.082**	0.047
		C	.5	**0.081**	**0.083**	**0.105**
			.7	**0.090**	**0.082**	**0.079**
			.9	**0.120**	**0.081**	0.046
		L	.5	**0.264**	**0.090**	**0.080**
			.7	**0.562**	**0.099**	**0.068**
			.9	**0.931**	**0.132**	0.050
	500/500	N	.5	**0.059**	**0.064**	**0.071**
			.7	**0.063**	**0.064**	**0.063**
			.9	**0.108**	**0.061**	0.051
		C	.5	**0.064**	**0.060**	**0.069**
			.7	**0.076**	**0.065**	**0.061**
			.9	**0.099**	**0.065**	0.057
		L	.5	**0.229**	**0.065**	**0.060**
			.7	**0.517**	**0.067**	**0.060**
			.9	**0.910**	**0.076**	0.050
Power (noninvariance)	200/200	N	.5	0.199	0.218	0.721
			.9	0.262	0.277	0.958
		C	.5	0.234	0.222	0.747
			.9	0.302	0.284	0.968
		L	.5	0.693	0.203	0.590
			.9	0.823	0.283	0.889
	500/500	N	.5	0.446	0.448	0.989
			.9	0.606	0.602	1.000
		C	.5	0.510	0.481	0.992
			.9	0.681	0.651	1.000
		L	.5	0.856	0.365	0.961
			.9	0.963	0.583	1.000

Note. LFM = linear factor model; CFM = categorical factor model; ML = maximum likelihood, MLR = robust maximum likelihood; WLSMV = weighted least squares mean and variance-adjusted. n_1/n_2 = sample size of Group 1/Group 2. τ = distribution of responses (N = approximately normal, C = censored, L = L-shaped). λ = factor loading for all 10 invariant items under the full metric invariance condition or for the 2 noninvariant items in the focal group under the metric noninvariance condition (where the loadings for the other items were .7). For the Type I error rates, values are boldfaced if they exceed .0415, .0585, (i.e., if they are more than 2 standard errors away from .05).

conditions, where inflation was greater for the small sample size and less skewed conditions.

In addition to size, the Monte Carlo estimated power of the chi-square test of the exact-fit hypothesis was evaluated by examining the average rejection rate for each

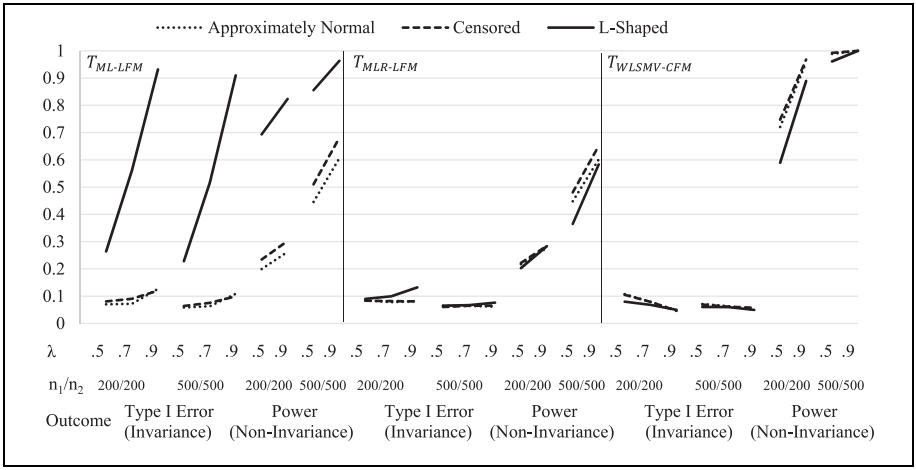

Figure 1. Monte Carlo estimated Type I error rate and power of the chi-square test of the exact-fit hypothesis for evaluating metric invariance.

Note. LFM = linear factor model; CFM = categorical factor model. λ = factor loading for all 10 invariant items under the full metric invariance condition or for the 2 noninvariant items in the focal group under the metric noninvariance condition (where the loadings for the other items were .7). n_1/n_2 = sample size of Group 1/Group 2.

of the two noninvariant conditions. Because the error rate was inflated under most conditions, power must be interpreted with caution. In particular, it is meaningless to evaluate power for the linear model with ML estimation under the L-shaped distribution. Nonetheless, focusing on conditions in which the Type I error rate was below .10, certain comparisons are worth noting. As expected, across conditions, power increased as a function of increased sample size and increased magnitude of metric noninvariance. In comparing model–estimator combinations, T_{ML-LFM} and $T_{MLR-LFM}$ demonstrated similar levels of power, while $T_{WLSMV-CFM}$ demonstrated much greater levels of power. For $T_{MLR-LFM}$ and $T_{WLSMV-CFM}$, power was noticeably lower for the most skewed (L-shaped) distribution condition.

Change in Model Fit

See Table 3 and Figure 2 for the Monte Carlo estimated Type I error rate and power of the chi-square and likelihood ratio difference tests for the equal-fit hypothesis across generation and analysis conditions.

Linear Factor Model. Compared to T_{ML-LFM}, ΔT_{ML-LFM} better held the correct size, particularly when the distribution of responses was approximately normal. The Type I error rate was not affected by sample size, but for the censored and L-shaped distributions, the error rate decreased with increased magnitude of the factor loadings.

Table 3. Monte Carlo Estimated Type I Error Rate and Power of the Chi-Square and Likelihood Ratio Difference Tests of the Equal-Fit Hypothesis for Evaluating Metric Invariance.

				LFM		TFM		CFM		
Outcome	n_1/n_2	τ	λ	ML	MLR	ML	MLR	ML	MLR	WLSMV
Type I error rate (invariance)	200/200	N	.5	0.052	0.051	**0.075**	0.052	0.058	0.053	**0.097**
			.7	0.054	0.056	**0.104**	**0.059**	0.044	0.055	**0.088**
			.9	0.058	0.055	**0.155**	0.058	**0.038**	**0.070**	**0.078**
		C	.5	**0.041**	0.050	**0.061**	0.048	0.057	0.049	**0.091**
			.7	**0.021**	0.048	**0.070**	0.053	0.044	0.057	**0.085**
			.9	**0.016**	0.052	**0.096**	0.054	**0.034**	**0.064**	**0.077**
		L	.5	**0.077**	0.044	**0.027**	0.046	0.056	0.044	**0.077**
			.7	0.057	0.047	**0.011**	0.053	0.046	0.053	**0.075**
			.9	**0.032**	0.051	**0.008**	0.051	**0.034**	**0.062**	**0.072**
	500/500	N	.5	0.053	0.053	**0.076**	0.053	0.054	0.050	**0.066**
			.7	0.051	0.052	**0.098**	0.050	**0.041**	0.055	**0.065**
			.9	0.055	0.051	**0.144**	0.049	**0.029**	0.054	0.057
		C	.5	**0.041**	0.053	**0.060**	0.052	0.055	0.053	**0.069**
			.7	**0.026**	0.053	**0.069**	0.052	**0.039**	0.050	**0.063**
			.9	**0.016**	0.054	**0.098**	0.053	**0.032**	0.058	**0.062**
		L	.5	**0.069**	0.048	**0.026**	0.049	0.052	0.049	**0.061**
			.7	0.055	0.050	**0.012**	0.051	0.043	0.051	**0.060**
			.9	**0.031**	0.054	**0.007**	0.054	**0.029**	0.054	**0.059**
Power (noninvariance)	200/200	N	.5	0.450	0.454	0.506	0.395	0.660	0.685	0.804
			.9	0.591	0.590	0.629	0.500	0.989	0.992	0.981
		C	.5	0.408	0.524	0.439	0.393	0.679	0.702	0.818
			.9	0.580	0.700	0.546	0.485	0.994	0.995	0.987
		L	.5	0.461	0.424	0.184	0.366	0.548	0.565	0.674
			.9	0.693	0.668	0.211	0.428	0.974	0.977	0.939
	500/500	N	.5	0.894	0.895	0.898	0.837	0.985	0.988	0.996
			.9	0.963	0.962	0.958	0.919	1.000	1.000	1.000
		C	.5	0.905	0.945	0.869	0.848	0.990	0.992	0.998
			.9	0.977	0.990	0.935	0.916	1.000	1.000	1.000
		L	.5	0.900	0.889	0.640	0.821	0.962	0.967	0.982
			.9	0.988	0.987	0.701	0.875	1.000	1.000	1.000

Note. LFM = linear factor model; TFM = Tobit factor model; CFM = categorical factor model; ML = maximum likelihood; MLR = robust maximum likelihood; WLSMV = weighted least squares mean and variance-adjusted. n_1/n_2 = sample size of Group 1/Group 2. τ = distribution of responses (N = approximately normal, C = censored, L = L-shaped). λ = factor loading for all 10 invariant items under the full metric invariance condition or for the 2 noninvariant items in the focal group under the metric noninvariance condition (where the loadings for the other items were .7). For the Type I error rates, values are boldfaced if they exceed .0415, .0585 (i.e., if they are more than 2 standard errors away from .05).

$\Delta T_{MLR-LFM}$ was even more stable, as it held the correct size across all generation conditions.

Tobit Factor Model. ΔG^2_{ML-TFM} was sensitive to the distribution of responses and magnitude of factor loadings but not to the sample size. The test was too liberal for the

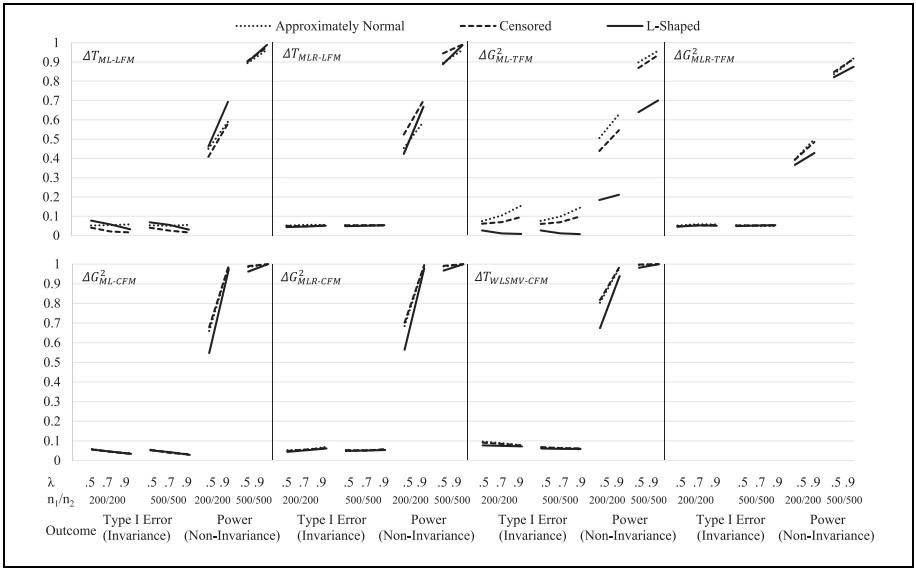

Figure 2. Monte Carlo estimated Type I error rate and power of the chi-square and likelihood ratio difference tests of the equal-fit hypothesis for evaluating metric invariance.
Note. ML = maximum likelihood; MLR = robust maximum likelihood; WLSMV = weighted least squares mean and variance-adjusted; LFM = linear factor model; TFM = Tobit factor model; CFM = categorical factor model. λ = factor loading for all 10 invariant items under the full metric invariance condition or for the 2 noninvariant items in the focal group under the metric noninvariance condition (where the loadings for the other items were .7). n_1/n_2 = sample size of Group 1/Group 2.

approximately normal and censored response distributions and too conservative for the L-shaped response distribution, where these patterns became even more apparent with increased magnitude of the factor loadings. $\Delta G^2_{MLR-TFM}$ performed much better, holding the correct size for all but one of the generation conditions. This condition corresponded to the approximately normal response distribution, where a TFM would not generally be applied.

Categorical Factor Model. ΔG^2_{ML-CFM} and $\Delta G^2_{MLR-CFM}$ performed relatively well, although ΔG^2_{ML-CFM} was too conservative when the magnitude of the factor loadings was large, and $\Delta G^2_{MLR-CFM}$ was too liberal when the magnitude of the factor loadings was large and the sample size was small. $\Delta T_{WLSMV-CFM}$, on the other hand, was consistently too liberal, with greater inflation occurring for the small sample size and less skewed (approximately normal and censored) response distribution conditions.

The power of the difference tests was examined by considering the noninvariant conditions. Again, power must be interpreted with caution due to instances in which the tests did not hold the correct size—only general patterns should be considered. As before, power increased with increased sample size and magnitude of

noninvariance. Across conditions, power levels were generally similar across estimators for a given measurement model, although $\Delta T_{WLSMV-CFM}$ was more powerful than G^2_{ML-CFM} and $\Delta G^2_{MLR-CFM}$ when the magnitude of noninvariance was small. In contrast, power levels were noticeable different across each of the three measurement models. Specifically, power was greatest for the CFM tests, and smallest for the Tobit tests. In general, power was greater for the less skewed response distribution conditions.

Discussion

We evaluated the impact of several model parameterization and estimation methods on the performance of the chi-square test of the exact-fit hypothesis and chi-square and likelihood ratio difference tests of the equal-fit hypothesis in the context of evaluating MI with approximately continuous ordered polytomous data. Our study makes several novel contributions to the MI literature by providing (1) an evaluation of understudied test statistics (i.e., $T_{MLR-LFM}$, ΔG^2_{ML-TFM}, $\Delta G^2_{MLR-TFM}$, and $\Delta G^2_{MLR-CFM}$), (2) a more elaborate factorial comparison of the various model parameterization and estimation methods, considering both Type I error rates *and* power, and (3) an evaluation of additional study factors such as the magnitude of the factor loadings.

With respect to evaluating overall model fit, T_{ML-LFM} was extremely unstable. In line with past research, but now extended to the multiple-group case, the Type I error rate was particularly inflated when the observed responses were highly skewed (cf., Babakus et al., 1987; B. Muthén & Kaplan, 1985). We also found that T_{ML-LFM} was sensitive to the magnitude of factor loadings. This finding provides greater support for past research that relied on a small number of replications (cf., Lubke & Muthén, 2004).

$T_{MLR-LFM}$ and $T_{WLSMV-CFM}$ were clearly preferred to T_{ML-LFM} as they demonstrated more controlled Type I error rates, although the error rate for $T_{MLR-LFM}$ was inflated when the sample size was small, and the error rate of $T_{WLSMV-CFM}$ was inflated when both the sample size was small (as observed within a single-group context; Flora & Curran, 2004) and the magnitude of factor loadings was large. The primary distinction between $T_{MLR-LFM}$ and $T_{WLSMV-CFM}$ was in their power to detect model misspecification. Comparatively speaking, $T_{MLR-LFM}$ was extremely underpowered. This may be particularly problematic for evaluations of MI in high-stakes settings where issues of fairness are at a forefront. In these settings, it may be better to error on the side of overflagging noninvariance in order to allow subject matter experts to perform further investigations. Thus, at least under the conditions examined in our study, $T_{WLSMV-CFM}$ may be preferred to $T_{MLR-LFM}$ (with the caveat that the chance of making a Type I error may be slightly elevated).

In addition to evaluating overall fit of the metric invariance model, change in fit between the configural and metric invariance models was examined. In terms of the Type I error rate, $\Delta T_{MLR-LFM}$, $\Delta G^2_{MLR-TFM}$, ΔG^2_{ML-CFM}, and $\Delta G^2_{MLR-CFM}$ generally

performed comparably and adequately. Although ΔT_{ML-LFM} performed considerably better than its exact-fit counterpart (T_{ML-LFM}), its performance fluctuated across the different observed response distributions and magnitudes of factor loadings, as did the performance of ΔG^2_{ML-TFM}. Consistent with the findings of Sass et al. (2014), $\Delta T_{WLSMV-CFM}$ was too liberal when the sample size was small. In line with Kim and Yoon (2011), but now extended to use of a free-baseline approach, the full-information CFM approaches (ΔG^2_{ML-CFM} and $\Delta G^2_{MLR-CFM}$) generally outperformed the limited-information CFM approach ($\Delta T_{WLSMV-CFM}$).

In comparing the tests that held the correct size, ΔG^2_{ML-CFM} and $\Delta G^2_{MLR-CFM}$ demonstrated the greatest power to detect metric noninvariance, and $\Delta G^2_{MLR-TFM}$ demonstrated the least power. This lack of power suggests that the TFM may not be particularly useful in the context of evaluating MI with ordered polytomous data, at least not under the conditions examined in this study. On the other hand, the full-information CFM approaches appear to perform well. For particularly complex models in which numerical integration is not a viable option, a limited-information CFM approach might be considered if the sample size is not too small. Although $\Delta T_{WLSMV-CFM}$ was found to be too liberal in many instances, the Type I error rate never exceeded .10. When taking into account the considerable gains in power achieved by using $\Delta T_{WLSMV-CFM}$ over $\Delta G^2_{MLR-LFM}$ and $\Delta G^2_{MLR-TFM}$, minor inflations in the Type I error rate may be of less concern. As we previously noted, power is a particularly important consideration when assessing MI in high-stakes contexts.

As with any simulation study, we considered only a finite number of factors that may influence the performance of the test statistics. In line with previous studies (e.g., French & Finch, 2006; Meade & Bauer, 2007; Yoon & Millsap, 2007), we focused on metric invariance. However, past research has shown that the type of measurement noninvariance (metric, scalar, or both) affects the relative performance of the test statistics across different measurement specifications and model estimators (Kim & Yoon, 2011; Sass et al., 2014), as does the presence of structural noninvariance (e.g., group mean differences; Lubke & Muthén, 2004). Although the CFM approaches were more powerful than the linear approaches for detecting metric noninvariance, the linear approaches may be more powerful for detecting scalar noninvariance (cf., Sass et al., 2014). Likewise, it is possible that other estimator and measurement model combinations may perform better than the ones evaluated in our study. In particular, Bayesian estimation may be useful for evaluating MI when dealing with complex models and small sample sizes (Sinharay, 2013).

Another limitation of our study is that we focused only on global assessments of model fit and change in model fit. In high-stakes settings, examinations are typically done at the item level (e.g., Kim & Yoon, 2011). Relatedly, we examined only the performance of test statistics; we did not compare model-estimator combinations with respect to parameter recovery or variance estimation. Point estimation is particularly important when using an effect size paradigm to assess MI (Wang et al., 2013).

Finally, although we used past research on MI and recommendations from the broader literature on latent variable modeling to guide our generating model and

simulation conditions, our generating parameters were not based on estimates from actual data. Whereas simplifications like the assumption of tau equivalence allowed us to minimize extraneous variability across conditions (thereby increasing the internal validity of our study), such simplifications also weakened the external validity of our study. In particular, the results of our study may not generalize to other contexts. Of course, this is a limitation of any study, even when parameters are based on actual data. Because it is impossible to investigate all possible scenarios, we encourage researchers to use the Monte Carlo facilities available in M*plus* (see B. Muthén, 2002) or other software environments to evaluate the impact of measurement parameterization and estimation decisions under data conditions that are directly relevant to their study.

The results of our study have important implications for researchers in the fields of psychology and education where the use of Likert-type scales is widespread and the need to investigate MI is essential. In line with Hayduk et al. (2007) and Kline (2016), we believe that researchers should always report *and* give serious consideration to the tests of the exact-fit and equal-fit hypotheses. All too often, researchers gloss over evidence of misfit in the form of significant chi-square tests, as if they feel compelled to find statistical support for the underlying hypothesis of MI. However, detection of noninvariance is crucial for maintaining fair testing procedures and ensuring validity of group comparisons. Although past research has identified limitations with these test statistics when applied to Likert-type data, our study demonstrates that selecting an appropriate estimation method and model parameterization can help mitigate such limitations. Furthermore, examination of global test statistics is only the first step in evaluating MI. A significant result merely alerts researchers to a potential problem. Regardless of whether the hypotheses of exact-fit and equal-fit are rejected, researchers must inspect local fit in the form of model residuals. If residuals are small and unsystematic, then MI may still be practically supported. As Kline (2016) reminds us, ''At the end of the day, regardless of whether or not you have retained a model, the real honor comes from following to the best of your ability a thorough testing process to its logical end'' (p. 269).

Authors' Note

Any errors or omissions are solely the responsibility of the authors. The opinions expressed herein are those of the authors and should not be considered reflective of the funding agency.

Acknowledgments

The authors thank R. J. de Ayala and Lesa Hoffman for feedback on an earlier draft.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Preparation of this article was supported by a grant awarded to Susan M. Sheridan and colleagues (IES No. R305C090022) by the Institute of Education Sciences.

Notes

1. Other models for ordered polytomous data include the rating scale (Andrich, 1978) and partial credit (Masters, 1982) models that assume all items are equally discriminating, and the generalized partial credit model (Muraki, 1992) that, like the GRM, does not make this assumption. We focus on the GRM because it is the default parameterization in M*plus*, a software program that is often used to assess measurement invariance.
2. See Casella and Berger (2002, p. 385) for further definition of a size α test.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573. doi:10.1007/BF02293814

Asparouhov, T., & Muthén, B. (2006). *Robust chi square difference testing with mean and variance adjusted test statistics* (M*plus* Web Notes No. 10). Los Angeles, CA: Muthén & Muthén.

Babakus, E., Ferguson, C. E., Jr., & Jöreskog, K. G. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. *Journal of Marketing Research*, *24*, 222-228. doi:10.2307/3151512

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, *42*, 815-824. doi:10.1016/j.paid.2006.09.018

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*, 186-203. doi:10.1207/s15328007sem1302_2

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Brown, T. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Wadsworth.

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

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford.

Fan, X., & Sivo, S. A. (2009). Using Δgoodness-of-fit indexes in assessing mean structure invariance. *Structural Equation Modeling*, *16*, 54-69. doi:10.1080/10705510802561311

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

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods*, *14*, 275-299. doi: 10.1037/a0015825

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

Hayduk, L. Cummings, G. G., Boadu, K., Pazderka-Robinson, H., & Boulianne, S. (2007). Testing! testing! One, two three: Testing the theory in structural equation models! *Personality and Individual Differences*, *42*, 841-850. doi:10.1016/j.paid.2006.10.001

IRT in Mplus. (2013). *IRT in Mplus* (M*plus* Technical Appendix). Los Angeles, CA: Muthén & Muthén.

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

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford.

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology*, *4*, 73-79. doi: 10.1027/1614-2241.4.2.73

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*, 514-534. doi:10.1207/s15328007sem1104_2

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174. doi:10.1007/BF02296272

Maydeu-Olivares, A., & Cai, L. (2006). A cautionary note on using G^2(dif) to assess relative model fit in categorical data analysis. *Multivariate Behavioral Research*, *41*, 55-64. doi: 10.1207/s15327906mbr4101_4

McBee, M. (2010). Modeling outcomes with floor or ceiling effects: An introduction to the Tobit model. *Gifted Child Quarterly*, *54*, 314-320. doi:10.1177/0016986210379095

Meade, A. W., & Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling*, *14*, 611-635. doi: 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. doi:10.1037/0021-9010.93.3.568

Meredith, W., & Horn, J. (2001). The role of factorial invariance in modeling growth and change. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 203-240). Washington, DC: American Psychological Association.

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.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-176. doi:10.1177/014662169201600206

Muthén, B. (2002). *Using Mplus Monte Carlo simulations in practice: A note on assessing estimation quality and power in latent variable models* (M*plus* Web Notes No. 1). Los Angeles, CA: Muthén & Muthén.

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, *38*, 171-189. doi:10.1111/j.2044-8317.1985.tb00832.x

Muthén, B. O., du Toit, S. H. C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes* (M*plus* Article No. 75). Los Angeles, CA: Muthén & Muthén.

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

Narayanan, P. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, *20*, 257-274. doi:10.1177/014662169602000306

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society.

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. doi:10.1080/10705511.2014.882658

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*, 507-514. doi:10.1007/BF02296192

Sinharay, S. (2013). An extension of a Bayesian approach to detect differential item functioning. *Journal of Applied Measurement*, *14*, 149-158.

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

Suh, Y., & Cho, S.-J. (2014). Chi-square difference tests for detecting differential functioning in a multidimensional IRT model: A Monte Carlo Study. *Applied Psychological Measurement*, *38*, 359-375. doi:10.1177/0146621614523116

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, *26*, 24-36. doi:10.2307/1907382

Wang, W., Tay, L., & Drasgow, F. (2013). Detecting differential item functioning of polytomous items for an ideal point response process. *Applied Psychological Measurement*, *37*, 316-335. doi:10.1177/0146621613476156

Woods, C. M. (2011). DIF testing for ordinal items with Poly-SIBTEST, the Mantel and GMH tests, and IRT-LR-DIF when the latent distribution is nonnormal for both groups. *Applied Psychological Measurement*, *35*, 145-164. doi:10.1177/0146621610377450

Yoon, M., & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling*, *14*, 435-463. doi:10.1080/10705510701301677

Yuan, K. H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. In M. E. Sobel & M. P. Becker (Eds.), *Sociological methodology 2000* (pp. 165-200). Washington, DC: American Sociological Association.