Educational and Psychological Measurement
2016, Vol. 76(4) 533–561
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0013164415603764
epm.sagepub.com
SAGE

The Role of Measurement Quality on Practical Guidelines for Assessing Measurement and Structural Invariance

Yoonjeong Kang[1], Daniel M. McNeish[2], and Gregory R. Hancock[2]

Abstract

Although differences in goodness-of-fit indices (ΔGOFs) have been advocated for assessing measurement invariance, studies that advanced recommended differential cutoffs for adjudicating invariance actually utilized a very limited range of values representing the quality of indicator variables (i.e., magnitude of loadings). Because quality of measurement has been found to be relevant in the context of assessing data-model fit in single-group models, this study used simulation and population analysis methods to examine the extent to which quality of measurement affects ΔGOFs for tests of invariance in multiple group models. Results show that ΔMcDonald's NCI is minimally affected by loading magnitude and sample size when testing invariance in the measurement model, while differences in comparative fit index varies widely when testing both measurement and structural variance as measurement quality changes, making it difficult to pinpoint a common value that suggests reasonable invariance.

Keywords

invariance testing, fit index, factor indicator reliability, multiple group model, structural equation modeling

[1]American Institutes for Research, Washington, DC, USA
[2]University of Maryland, College Park, MD, USA

Corresponding Author:
Daniel M. McNeish, University of Maryland, 1230 Benjamin Building, College Park, MD 20742-1115, USA.
Email: dmcneish@umd.edu

In social and behavioral studies, researchers are often interested in comparing groups on latent constructs that are not observed directly. Although research interests frequently concern differences in an underlying structural model of the latent variables across key groups such as sex, ethnicity, nationality, or socioeconomic status (e.g., Fonseca-Pedrero et al., 2010; Moura, dos Santos, Rocha, & Matos, 2010), in order to make group comparisons of those latent constructs' relations meaningful it is prerequisite to ensure that constructs are indeed the same in all populations (Horn & McArdle, 1992; Meredith & Teresi, 2006). If different latent constructs are captured by a measurement instrument in different populations, group comparisons involving the latent constructs could be meaningless and invalid. As Vandenberg and Lance (2000) stated, if a set of items does not mean the same thing to different groups, group comparison on the latent constructs ''may be tantamount to comparing apples and spark plugs'' (p. 9).

Construct equivalence is a conceptual notion and is related to theoretical validity (van de Vijver & Tanzer, 2004). Thus, construct equivalence cannot be statistically tested and often rests on substantive theories or strong beliefs by researchers. Despite this, one statistical procedure, measurement invariance testing, has been used to collect evidence of construct equivalence, that being measurement equivalence. Even though measurement equivalence may not be a sufficient condition for ensuring construct equivalence, establishing measurement equivalence is often viewed as a critical assumption when comparing latent constructs and their relations across populations (e.g., Cheung & Rensvold, 1999; Meredith, 1993; Meredith & Teresi, 2006).

Measurement Invariance

Types of Invariance

Tests of measurement invariance across populations are typically conducted through multigroup confirmatory factor analysis (MCFA) by examining the extent to which model parameters are invariant. Specifically, Meredith (1993) and Meredith and Teresi (2006) described three types of measurement invariance: weak, strong, and strict factorial invariance.[1] The first, and least restrictive, is weak factorial invariance, which is also referred to as factor loading invariance (or metric invariance). Weak factorial invariance indicates that measured indicator variables (e.g., items) are related to their latent constructs in the same way across the populations of interest; however, factor means, factor variances and covariances, intercepts, and error variances do not necessarily have to be equal across population to achieve weak factorial invariance. If weak factorial invariance does not hold, the equivalence of the latent constructs across populations comes into question, and hence group comparisons might become tenuous. Often, achieving weak factorial invariance is considered sufficient to proceed with group comparisons for the structural covariance model (Byrne, Shavelson, & Muthén, 1989; Horn, McArdle, & Mason, 1983).[2]

Strong factorial invariance indicates that the factor loadings and intercepts of measured variables are the same across populations. The factor means, factor variances and covariances, and error variances do not necessarily need to be invariant across populations in order to achieve strong factorial invariance. The intercept invariance indicates that any differences in means on the measured variables correspond to differences in means on the latent constructs alone, and hence strong factorial invariance allows a researcher to compare group differences on intercepts of the latent constructs directly (Brown, 2006).

Finally, with strict factorial invariance, variances and covariances of errors, as well as factor loadings and intercepts of measures, are constrained across populations. Under strict factorial invariance, any differences in variances and covariance of measured variables are attributable to differences in variances and covariances of latent constructs, and thus it allows unbiased group comparisons on the latent constructs.

Given that these three types of invariance require that particular parameters (e.g., factor loadings with weak invariance, factor loadings and indicator intercepts with strong invariance) are constrained in all populations, strictly speaking, it is unlikely for these equalities to hold exactly in reality (Little, Card, Slegers, & Ledford, 2007; Millsap & Meredith, 2004). Some researchers have proposed partial measurement invariance in which a subset of parameters is invariant while another subset of parameters may not be invariant and are allowed to be freely estimated across groups (Byrne et al., 1989). When only partial measurement invariance is upheld, however, it is debatable whether comparisons of latent construct across populations are meaningful because partial measurement invariance does not support measurement equivalence across populations, which may be a necessary condition for ensuring construct equivalence. Furthermore, some researchers have suggested that partial measurement invariance may indicate construct nonequivalence because nonequivalence of factor loading means that the latent constructs are inferred from the observed variables in a different way across populations (e.g., Meredith & Teresi, 2006). Therefore, some researchers have argued that weak factorial invariance is the minimum needed to consider construct equivalence (Cheung & Rensvold, 1999; Little et al., 1997; Meredith & Teresi, 2006).

Assessing Measurement Invariance

In order to test each level of measurement invariance, two measurement models (e.g., an unconstrained model and a model with factor loadings, intercepts, and/or error variances constrained across groups) generally are fit to the same sample data and compared. One common statistic used to compare the unconstrained and constrained models is the difference between the minimum fit function chi-square statistics across models (ΔT_{ML}; Jöreskog, 1971). To calculate ΔT_{ML}, first the model is fit without any constraints imposed between the groups, and the T_{ML} statistic is calculated by

$$T_{ML} = (n-1)\min\left(\ln|\mathbf{\Sigma}| - \ln|\mathbf{S}| + \mathrm{tr}\left[\mathbf{S}\mathbf{\Sigma}^{-1}\right] - p\right),$$

where $\mathbf{S}$ is the observed covariance matrix, $\mathbf{\Sigma}$ is the model-implied covariance matrix, p is the number of observed variables, and n is sample size. Then, the constraints are imposed on the model (e.g., constraining factor loadings to be equal if testing weak factorial invariance) and the T_{ML} statistic is again calculated for the constrained model. The constrained model should fit the data worse or at best equally well (i.e., $T_{\text{unconstrained}} \leq T_{\text{constrained}}$) because the data are being modeled with fewer parameters (e.g., one estimated loading applies to both groups rather than each group having a separate loading estimate). Under standard assumed conditions, the difference between $T_{\text{constrained}}$ and $T_{\text{unconstrained}}$ is chi-square distributed with the degrees of freedom equal to the difference in the degrees of freedom of the two models (which is also equal to the number of constrained parameters). The difference in the T_{ML} statistics for the constrained and unconstrained models can be used to test whether the fit for the constrained model is significantly worse than the unconstrained model. A statistically significant ΔT_{ML} test suggests noninvariance for the constrained parameters and that certain parameters should to be freely estimated for data-model fit to improve (i.e., implying partial invariance or noninvariance).

Although ΔT_{ML} is most frequently used, there has been some opposition to using it to assess invariance. For instance, it has been found to be highly sensitive to sample size in invariance testing because T_{ML} for the unconstrained and unconstrained models becomes overpowered with larger sample sizes and trivial differences between groups may be flagged as noninvariant across populations (Cheung & Rensvold, 2002; Meade, Johnson, & Braddy, 2008; Wu, Li, & Zumbo, 2008). For instance, Brown (2006) noted that with larger sample sizes it is possible for the ΔT_{ML} test to be significant but that follow-up diagnostics do not detect that unconstraining any of the parameters would lead to appreciably better data-model fit (see Campbell-Sills, Liverant, & Brown, 2004, for an example of this in an applied study). Similar types of arguments were, in fact, the foundation for developing and favoring goodness of fit indices (e.g., root mean square error of approximation [RMSEA], standardized root mean square residual [SRMR], comparative fit index [CFI]) in single group models. However, ΔT_{ML} is more commonly reported in applied multiple group studies compared to the reporting of T_{ML} in single group studies. Cheung and Rensvold (2000) and Vandenberg and Lance (2000) have noted this apparent double standard between single group and multiple group tests: If T_{ML} is not trusted for assessing fit in single group models, then why should it be trusted for multiple group models with the same data?

For this reason, research has suggested to use alternative goodness-of-fit indices (GOFs) in invariance testing that are potentially less sensitive to sample size. Cheung and Rensvold (2002) discussed ΔGOFs, which take the difference of a GOF for an unconstrained model and a constrained model in the same vein as ΔT_{ML}. This approach has received support in the literature, such as Byrne (2008) stating,

> Researchers have argued that this $\Delta\chi^2$ value is an impractical and unrealistic criterion upon which to base evidence of equivalence. Thus, there has been a trend towards basing

> comparative models on the difference between the CFI values as a more practical approach to determining the extent to which models are equivalent. (p. 878)

Cheung and Rensvold (2002) outlined four desirable properties of ΔGOFs used for testing measurement invariance:

1. ΔGOFs should not be sensitive to the overall fit in the baseline model.
2. ΔGOFs should not be sensitive to model complexity.
3. ΔGOFs should not be redundant with other GOFs.
4. ΔGOF should not be sensitive to sample size.

Following these four criteria, Cheung and Rensvold (2002) examined 20 GOFs. Their simulations found that ΔCFI, ΔGamma-hat, ΔMcDonald's NCI (hereafter abbreviated ΔMNCI), ΔIFI, and ΔRNI adhere to the desirable properties. Due to high correlation among ΔIFI, ΔCFI, and ΔRNI (and thus violating the third desirable property), they suggested reporting only one of these three indices. Given that CFI is a popular index in confirmatory factor analysis, they recommended using ΔCFI to assess measurement invariance along with ΔGamma-hat and ΔMNCI, which also performed well in accordance with the four desirable properties. Furthermore, Cheung and Rensvold provided empirically derived cutoff values for ΔCFI, ΔGamma-hat, and ΔMNCI that were −0.01, −0.001, and −0.02,[3] respectively, which approximately correspond to the empirical 1st percentile across all types of invariance tests including invariance tests at both the measurement and structural levels.

F. F. Chen (2007) extended this work by further examining the performance of ΔGOFs specifically for detecting measurement noninvariance, focusing on five ΔGOFs, ΔCFI, ΔRMSEA, ΔSRMR, ΔGamma-hat, and ΔMNCI, under both measurement invariance and noninvariance conditions. With one-factor models, she found that all ΔGOFs except for ΔSRMR were unaffected by sample size and three types invariance (weak, strong, strict), resulting in common cutoff values for those four ΔGOFs across invariance tests. She provided empirically derived cutoff values for ΔCFI, ΔRMSEA, ΔGamma-hat, and ΔMNCI that were −0.005, 0.01, −0.005, and −0.01, respectively, at $\alpha = .01$ across three types of invariance. For ΔSRMR, she provided cutoffs of 0.025 for weak factorial invariance and 0.005 for strong or strict factorial invariance. Inconsistent with Cheung and Rensvold (2002), she found that ΔGamma-hat had a strong relation with ΔCFI and thus did not recommend using ΔGamma-hat. Moreover, she did not recommend using ΔMNCI because ΔRMSEA and ΔSRMR had slightly higher power than ΔMNCI for detecting measurement noninvariance.

A similar study by Meade et al. (2008) was conducted to empirically derive cutoff values for ΔGOFs. Similar to F. F. Chen (2007), they expanded on the work by Cheung and Rensvold (2002) by investigating the performance of 20 ΔGOFs under the both measurement invariance and noninvariance (partial invariance) conditions.

In general, the results were consistent with those of Cheung and Rensvold (2002), finding that ΔIFI, ΔRNI, ΔGamma-hat, ΔCFI, and ΔMNCI performed relatively well with respect to the original criteria by Cheung and Rensvold (2002) and had relatively high power to detect measurement noninvariance. Inconsistent with Cheung and Rensvold (2002), however, but consistent with F. F. Chen (2007), it was found that ΔGamma-hat was also highly correlated with ΔIFI, ΔRNI, and ΔCFI. In the end, Meade et al. (2008) recommended using one of the four indices among ΔGamma-hat, ΔIFI, ΔRNI, and ΔCFI along with ΔMNCI. In their second simulation, they derived empirical cutoff values of ΔCFI and ΔMNCI under different conditions varying sample size, number of indicators per factor, and number of factors while holding other model parameters constant (e.g., factors' reliability). Consistent with Cheung and Rensvold (2002), they found that model complexity (e.g., number of factors and number of indicators) and sample size had little effect on ΔCFI, and use of a common cutoff value for ΔCFI that was empirically derived performed well in detecting lack of measurement invariance. Therefore, they provided a common cutoff value for ΔCFI of -0.002 to assess either weak or strong factorial invariance. Of interesting note was that ΔMNCI appeared to have different levels of power for detecting lack of measurement invariance depending on model complexity even though the effect size from ANOVA showed little effect of model complexity on ΔMNCI. Given this, they provided different empirically derived cutoff values for ΔMNCI based on the number of factors and indicators.

Shortcomings of ΔGOF Strategy

Despite the support in the literature for using the ΔGOF approach to assess invariance, the broad utility of the method has not been definitively demonstrated and many works have advocated for further research. For instance, Wu et al. (2008) stated that ''more research like Cheung and Rensvold's is needed to validate their findings in other settings'' (p. 6). Kline (2011) also mentioned in his popular text that ''specifically, it is unknown whether this rule of thumb [ΔCFI $\leq$ 0.01] would generalize to other models or data sets not directly studied by Cheung and Rensvold (2002)'' (p. 254). Brown (2006) similarly declared ''although the authors [Cheung & Rensvold] proposed cutoffs for three fit statistics, the validity of these proposals awaits further research'' (p. 303). Thus, because ΔGOFs are indices and not statistics with typical distributions, cutoff values that are indicative of invariance or noninvariance are derived from simulated empirical distributions that may not be broadly generalizable to conditions not included in the study used to derive those values.

Such conditions include measurement quality. Specifically, although Cheung and Rensvold (2002), F. F. Chen (2007), and Meade et al. (2008) explicitly or implicitly mentioned the role of measurement quality or factor indicator quality (e.g., the magnitude of the factor loadings) as a potential factor in assessing measurement invariance, these studies did not examine how measurement quality affects ΔGOFs in invariance testing. For instance, Cheung and Rensvold (2002) manipulated the factor loadings using two patterns for each level of number of factor's indicators (e.g., 1:1:1 and 1:1.25:1.5 for 3 indicators), but their focus was on factor loading pattern (e.g., homogeneous vs. heterogeneous) and thus was not related to measurement quality. Similar to Cheung and Rensvold (2002), all factor loadings were equal across all models with same structures in Meade et al. (2008). For example, the four-factor models where each factor had four indicators had equal factor loadings across all simulation conditions and was not a manipulated condition. In F. F. Chen (2007), the factor loadings were set to 0.90 for all factor loadings except for reference variable (which was fixed to 1) across all conditions. Measurement quality does, however, play an important role in the calculation of GOFs, as will be extensively discussed in the subsequent sections.

The Impact of Measurement Quality

The Role of Measurement Quality in Previous Studies

In order to better understand the relation between measurement quality and ΔGOFs, the former must be defined. Measurement quality can be determined by both factor loading magnitude and number of factor indicators in a model. Collectively, these inform the stability or reliability of the factor, which has been found to be a significant determinant of convergence and accuracy of parameter estimates (Gagné & Hancock, 2006). Furthermore, it also has been found that factor reliability is a significant factor in precision and power of measurement invariance testing along with sample size (Meade & Bauer, 2007).

Although factor reliability as determined by factor loading magnitude and number of indicators has been known to be an important factor in measurement invariance testing, the previous studies investigating ΔGOFs to assess measurement invariance did not examine the effect of measurement quality, and instead the factor reliability was constrained to be equal or within a very small range across all simulation conditions. For example, with a measure of factor's reliability by Fornell and Larcker (1981) (discussed in greater detail below) the simulation in Cheung and Rensvold (2002) fixed the reliability of each factor to be 0.80 in all conditions, and Meade et al. (2008) also fixed the reliability of each factor within a range from 0.74 to 0.82. Based on factor loadings and residual variances, the factor reliability in the models used in F. F. Chen (2007) was at least 0.92 across simulation conditions.

These previous studies cannot be faulted for not including all possible values for factor loadings as the simulation would become unwieldy; nonetheless, substantive invariance studies routinely have factor reliabilities that fall well below the rather optimistic conditions used in these studies. For recent examples across fields such as business, education, psychology, and public health that use ΔGOFs with factor reliabilities below what previous methodological studies have covered, see, for example, Karcher and Sass (2010); Pan, Rowney, and Peterson (2012);Savickas and Porfeli (2012); and Zhang et al. (2011).

Interestingly, Meade et al. (2008) found that cutoff values for ΔCFI and ΔMNCI approach zero as the number of indicators per factor increased. Their results may suggest that models with large number of factor indicators tend to yield smaller values of ΔGOFs. Even though Meade et al. (2008) explained that the reasons of the discrepancy in cutoff values for ΔCFI provided between their study and Cheung and Rensvold (2002) might be attributable to difference in models' complexity and level of invariance tests, it may be also due to difference in measurement quality of the models used in the two studies. Given the fact that overall measurement model quality in F. F. Chen (2007) and Meade et al. (2008) was higher than in Cheung and Rensvold (2002), the model with high-quality latent constructs would have less sampling variability in ΔGOFs and hence one would expect smaller ΔGOFs. This finding also implies that the same cutoff values should not be applied to assess measurement invariance when measurement models with latent variables have different quality, making the generalizability of such cutoffs highly suspect.

Overview of Reliability Measures

Various measures of factor reliability have been proposed by several researchers. One popular measure of factor reliability by Fornell and Larcker (1981) can be expressed for a *k*-indicator factor as follows:

$$\rho = \frac{\left(\sum_{i=1}^{k} \lambda_i\right)^2}{\left(\sum_{i=1}^{k} \lambda_i\right)^2 + \sum_{i=1}^{k} \mathrm{var}(\varepsilon_i)},$$

where λ_i indicates the factor loading of indicator *i* and ε_i indicates the residual variance of indicator *i*. Given that this measure mainly focuses on assessing composite scales that are created by factor indicators rather than the expected stability of the modeled latent construct itself, Hancock and Mueller (2001) advocated a measure of maximal reliability (or construct reliability) as a more appropriate measure of reliability of the latent construct. Construct reliability can be considered as the extent to which the latent construct is reproducible from its own measured indicators (Hancock & Mueller, 2001), and their coefficient *H* can be expressed for a single *k*-indicator factor using its standardized loadings (a_i) as follows:

$$H = \frac{\sum_{i=1}^{k} \frac{a_i^2}{(1-a_i^2)}}{1 + \sum_{i=1}^{k} \frac{a_i^2}{(1-a_i^2)}}.$$

This index has been shown to be directly linked to power in mean and covariance structure models (e.g., Hancock, 2001; Hancock & French, 2013).

The Role of Measurement Quality on GOFs

The effect of factor reliability on GOFs has been previously documented in the context of single-group models by Hancock and Mueller (2011) and Saris, Satorra, and van der Veld (2009). In an illustrative study, Hancock and Mueller constructed a population covariance matrix based on a six-factor, longitudinal structural model with each factor having three manifest indicator variables that each had the same standardized factor loading value in the population. To induce misfit, the fitted model was specified to lack nonnull structural paths that were present in the population model so that the model-implied covariance matrix did not exactly replicate the population covariance matrix. Without altering the misspecified paths, the values of the standardized factor loadings were manipulated so that they took on values between 0.40 and 0.95 in increments of 0.05, thereby systematically altering the reliability of the latent factors. After fitting the misspecified model to the population covariance matrices resulting from different loading patterns, frequently employed GOFs (i.e., SRMR, RMSEA, goodness-of-fit index [GFI], adjusted goodness-of-fit index [AGFI], and CFI) were recorded and compared for the range of factor loading values. Using the values in the appendix of Hancock and Mueller (2011), Figure 1 reproduces their results for RMSEA, SRMR, AGFI, CFI, and GFI. In the left panel of Figure 1, the dashed horizontal lines represent the Hu and Bentler (1999) suggested cutoffs for acceptable data-model fit where values above the dashed lines indicate poor data-model fit and values below the dashed lines indicate acceptable data-model fit. In right panel of Figure 1, AGFI, CFI, and GFI values above the dashed line indicate acceptable data-model fit, whereas values below the dashed line indicate poor data-model fit.

Intuitively, one might expect that less reliable latent factors would result in poorer data-model fit. However, even though the model was misspecified and should have been recognized as such, for standardized factor loadings less than 0.70 the tracked indices suggested acceptable data-model fit based on criteria set forth by Hu and Bentler (1999) including an SRMR at or below 0.08, RMSEA at or below 0.06, and CFI, GFI, and AGI at or above 0.95. The seemingly contradictory finding of obtaining better data-model fit under poorer measurement conditions is resolved, of course, by understanding that worse measurement quality leads to reduced sensitivity to detect latent misspecifications. Notwithstanding, this result was especially troubling because it implies that model fit index cutoff values that are ubiquitously referenced in the latent variable model literature (such as those of Hu and Bentler) cannot be universally applied because they themselves were derived under specific (and limited) quality of measurement conditions. Hu and Bentler (1999), for example, had only considered standardized factor loading values between 0.70 and 0.80, which is precisely where the approximate fit indices perform well in Hancock and Mueller (2011). Thus, by extension and as addressed in the current work, the differing derived ΔGOF cutoff values among Cheung and Rensvold (2002), F. F. Chen (2007), and Meade et al. (2008) may similarly be attributed to the role of different measurement quality contexts.

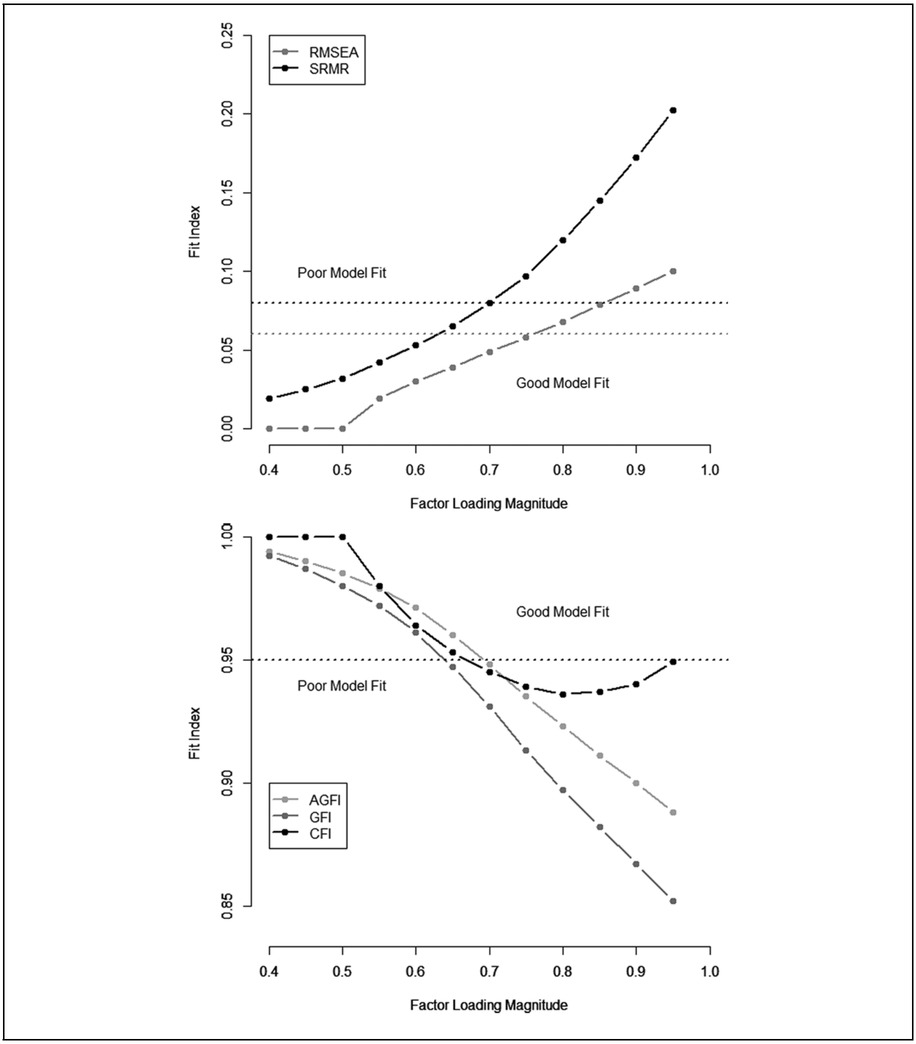

Figure 1. Replicated results from Hancock and Mueller (2011).

To elaborate, this study will explicate further the relation between measurement quality and ΔGOFs and will attempt to derive empirical cutoff values for ΔGOFs as a function of measurement model quality and sample size, thereby creating more refined practical guidelines for assessing measurement invariance. This study will explicitly focus on the effect that measurement quality has on ΔGOFs in assessing measurement invariance and the measurement quality has when assessing structural invariance. Additionally, the study will consider whether cutoffs can truly exist if they greatly vary based on quality of measurement.

Method

Measurement Invariance Study

As Cheung and Rensvold (2002), F. F. Chen (2007), and Meade et al. (2008) collectively advocated for ΔCFI and ΔMNCI, we will focus only on these two indices in the current study. Given that previous studies found little effect of number of factors in sampling variability in ΔCFI and ΔMNCI (Cheung & Rensvold, 2002; Meade et al., 2008), the current study also considered only a two-group, one-factor model in the simulation. Three conditions were manipulated in our simulation:

1. Number of indicators per factor (3 or 5 indicators per factor)
2. Factor loading magnitude (standardized values of 0.40 to 0.95 in increments of 0.05)
3. Sample size (100, 200, 300, 600, and 1,000 per group)

Even though previous studies found little effect for the number of factor indicators on ΔCFI and ΔMNCI, the number of indicators per factor was manipulated to thoroughly investigate the relation between measurement quality and the change in the two indices. Based on the factor loading magnitudes, construct reliability (*H*) ranged between 0.35 and 0.97 for the 3 indicator condition and between 0.49 and 0.98 for the 5 indicator condition. Even though two indices were previously found to be insensitive to sample size, sample size was manipulated to investigate whether there is an interaction effect between sample size and measurement quality. For each of the 120 conditions in the 12 × 5 × 2 design (12 factor loading conditions, 5 sample sizes conditions, and 2 number of indicators per factor conditions), we simulated 1,000 data sets, each drawn from a multivariate normal distribution. In addition, data were generated for the two groups such that they were invariant in terms of factor loadings (i.e., adhered to weak invariance).

For each condition, two models were estimated. In the first (unconstrained) model, model parameters were freely estimated across groups. In the second (constrained) model, factor loadings were constrained to be equal across groups to test weak factorial invariance. For model identification, the factor variances in both groups were set to 1 in the unconstrained model. For the constrained model, the factor variance in only one group was set to 1 while constraining all factor loadings to be equal across groups. The difference of CFI and MNCI between the unconstrained model and the constrained model was calculated over 1,000 replications. For each ΔCFI and ΔMNCI, the 5th and 1st percentiles of two indices were calculated to derive the empirical cutoff values at the $\alpha = .05$ and $\alpha = .01$ levels. The data generation and analysis of the sample data was conducted using the Mplus Version 6.0 (Muthén & Muthén, 2010), and SAS 9.2 was then used to perform additional analyses including ANOVA.

Structural Invariance Study

Although much of the previous literature on invariance testing is concerned with investigating the tenability of invariance in the measurement model (Finch & French, 2013), invariance of the structural model across groups can also be of interest substantively if reasonable evidence for measurement invariance exists. Similar methods can be applied to testing structural invariance as are applied to test measurement invariance such as ΔT_{ML} or ΔGOFs (Byrne, 2013). Moreover, substantive studies have applied the recommendations for ΔGOFs for measurement invariance from Cheung and Rensvold (2002) to structural invariance as well (see, e.g., H. Chen, Keith, Weiss, Zhu, & Li, 2010; Teo, Lee, Chai, & Wong, 2009), and these recommendations have been suggested for invariance testing broadly (Byrne, 2013).

Thus, the second study will investigate the role that measurement quality has on the ability of Cheung and Rensvold criteria for ΔCFI and ΔMNCI to detect nonnull differences in structural parameters in the presence of measurement invariance (i.e., if measurement quality affects Type II errors for structural parameters). Due to the more widespread methodological interest in assessing measurement invariance and due to the space limitations, the structural invariance study will use a population analysis using Hancock and Mueller (2011) as a template rather than a more expansive simulation study.

The population covariance matrix is created from a six-factor longitudinal design where three covarying exogenous factors each have a causal impact on one of three endogenous factors. All six factors were specified to have three manifest indicator variables each with identical population values for the factor loadings. All indicator variables were standardized to have a mean of 0 and variance of 1. The structure of the population model was identical in each of the two groups and the population values for the measurement model were identical across groups (i.e., weak factorial invariance) but the population values for the structural paths differed between groups. Figure 2 presents a conceptual path diagram for the structural and measurement models.

Similar to the measurement invariance study above and to Hancock and Mueller (2011), the population values for the factor loadings were manipulated from 0.40 to 0.95 in increments of 0.05. Between-group misfit was induced through the specification of different population values for the structural parameters across groups, while the analysis model constrained the structural paths to be equivalent between groups. The degree of misfit had three separate conditions. The first was a large discrepancy condition (Condition 1) where the structural paths for Group 1 were set to a standardized value of 0.80, while the structural paths for Group 2 were set to a standardized value of 0.10. Condition 2 was a medium discrepancy condition where the structural paths for Group 1 were set to a standardized value of 0.80, and the structural paths for Group 2 were set to a standardized value of 0.40. Condition 3 was a small discrepancy condition where the structural paths for Group 1 were set to a standardized value of 0.80 and the structural paths for Group 2 were set to a standardized value of 0.60.

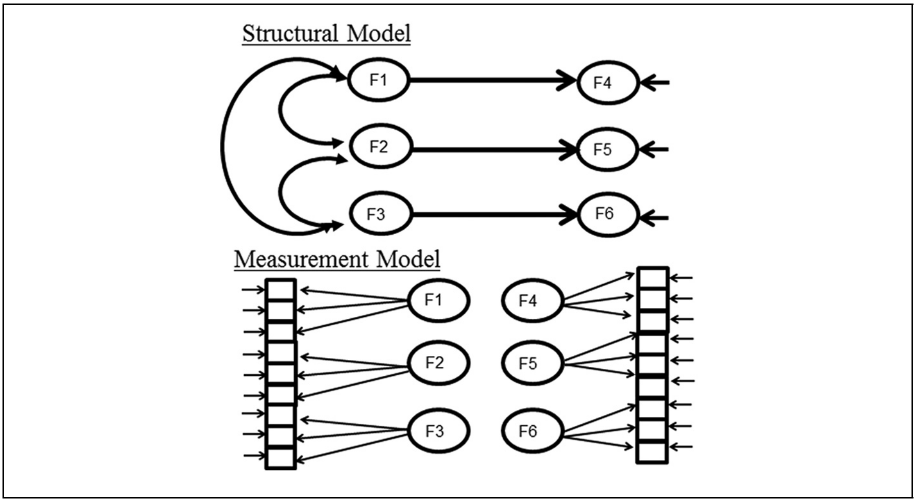

Figure 2. Pictorial depiction of population model for both groups.

For all three conditions, the exogenous factors had unit variance and covariances of 0.50 for both groups, although these parameters were not constrained to be equal in the analysis model. The latent variables were given scale by the reference variable method of setting a factor loading to 1 for each of the six factors. Since each indicator variable loaded identically for each factor and all indicator variables were standardized, the choice of which factor loading to constrain was arbitrary and thus the first variable was selected for each factor. Factor variances and indicator variable error variances were allowed to be freely estimated in both groups. Analyses were implemented using maximum likelihood in SAS 9.2 using Proc Calis because fit statistics and criteria are reported to four decimal places, which can better capture the nuances of ΔGOFs, which often have small magnitudes. A sample size of n =1,000 was used in each group.

Results

Measurement Invariance Study

Convergence Rates. Convergence rates for all conditions were found to be good, ranging from 96% to 100% across all conditions with two exceptions. The exceptions were observed in the two unconstrained models where the population factor loadings were 0.40 and 0.45, sample size was 100 per group, with three factor indicators. Specifically, the nonconvergence rates reached up to 23% and 10% when population factor loadings were 0.40 and 0.45, respectively. In the constrained models, convergence rates for all conditions were either 99% or 100% with one exception. The exception occurred in the model where the population factor loadings were 0.40, the sample size was 100 per each group, and three indicators per factor. However, the nonconvergence rate for that condition was still less than 6%.

Factor Loading Bias. In the conditions with nonconvergence rates exceeding 10%, the factor loading estimates were found to be biased. For the unconstrained model that exhibited nonconvergence rates as high as 23%, factor loading relative bias ranged between 9.7% and 21.8%. Given the ±5% criterion by Hoogland and Boomsma (1998) for negligibly biased estimates, the factor loading estimates under these conditions were considered to be unacceptable. Similar results also occurred under the condition where the population factor loadings were 0.45. In addition, unacceptable factor loading relative bias was observed in the baseline model where the sample size was 200 per group, the population factor loadings were 0.40, with three indicators per factor. These results were not surprising given that previous studies have consistently found that the combination of small sample sizes and low measurement quality have significant effects on recovery of parameter estimates (e.g., Gagné & Hancock, 2006). As expected, as sample size or factor loading magnitude increased, the convergence problem and magnitude of bias rapidly decreased. In most of conditions, the relative bias for the factor loading estimates was negligible based on criteria in Hoogland and Boomsma (1998), ranging from 0% to 4% in absolute value. Samples with convergence problems in the above three conditions with unacceptable factor loading relative bias were omitted from further analyses.

Measurement Quality. For the main analyses, the effects of sample size and factor loading magnitude on the ΔCFI and ΔMNCI were tested by conducting two-way analyses of variances (ANOVAs) for each index. Table 1 and 2 show the results of ANOVAs along with ω^2 effect size for ΔCFI and ΔMNCI when number of indicators per factor was 3 and 5, respectively. It was found that the sample size, factor loading magnitude, and sample size × factor loading magnitude interaction had significant effects on ΔCFI at the 0.01 level of significance regardless of the number of indicators per factor. When the number of indicators per factor was 3, the ω^2 effect sizes[4] of sample size, factor loading magnitude, and sample size × factor loading magnitude were 0.014, 0.030, and 0.013, respectively, for ΔCFI. When the number of factor indicators was increased to 5, the ω^2 effect sizes of sample size, factor loading magnitude, and sample size × factor loading magnitude decreased to 0.004, 0.013, and 0.008, respectively. These findings imply that sample size, factor loading magnitude, and sample size × factor loading magnitude have small, but noticeable, impacts on ΔCFI.

For ΔMNCI, the factor loading magnitude and sample size × factor loading magnitude interactions had significant effects at the 0.01 level of significance when the number of indicators per factor was 3. When the number of indicators per factor was increased to 5, however, only the sample size was found to have significant effect on ΔMNCI at the 0.01 level of significance. However, the ω^2 effect sizes showed that

Table 1. Results of the ANOVA Tests for ΔCFI and ΔMNCI (Number of Indicators = 3).

ΔCFI

Source	*df*	MS	*F*	ω^2
Between				
Sample size	4	0.11278	224.51	0.014
Factor loading	11	0.08826	175.69	0.030
Sample size × Factor loading	44	0.00967	19.24	0.013
Within	59,461	0.00050		
Total	59,520			

ΔMNCI

Source	*df*	MS	*F*	ω^2
Between				
Sample size	4	0.00001	1.00	<0.001
Factor loading	11	0.00005	7.13	<0.001
Sample size × Factor loading	44	0.00003	3.90	0.001
Within	59,461	0.00001		
Total	59,520			

Table 2. Results of the ANOVA Tests for ΔCFI and ΔMNCI (Number of Indicators = 5).

ΔCFI

Source	*df*	MS	*F*	ω^2
Between				
Sample size	4	0.00726	64.29	0.004
Factor loading	11	0.00811	71.89	0.013
Sample size × Factor loading	44	0.00139	12.36	0.008
Within	59,902	0.00011		
Total	59,961			

ΔMNCI

Source	*df*	MS	*F*	ω^2
Between				
Sample size	4	0.00013	9.09	<0.001
Factor loading	11	0.00002	1.66	<0.001
Sample size × Factor loading	44	0.00002	1.27	<0.001
Within	59,902	0.00001		
Total	59,961			

all effect sizes for the main effects and interactions effects were less than or equal to 0.001, indicating that there may be little practical impact on ΔMNCI.

Table 3 and Table 4 list the means, standard deviations, 5th percentile, and 1st percentile of ΔCFI and ΔMNCI for testing weak factorial invariance of each combination of sample size and factor loading magnitude when number of factor indicators was 3 and 5, respectively. As sample size and factor loading magnitude increased, the means and standard deviations of ΔCFI decreased, resulting in different 5th and 1st percentiles (i.e., cutoff values) of ΔCFI. As shown in Figure 3, models with higher factor loading magnitude (i.e., higher measurement quality) and bigger sample size tended to yield values of the 1st percentile of ΔCFIs with smaller magnitudes and vice versa. Graphs for the 5th percentile of ΔCFIs are similar and thus not presented. As expected, as shown in Figure 4, the number of indicators generally seemed to have little impact on the 1st percentiles of ΔCFI, which is consistent with previous studies (F. F. Chen, 2007; Cheung & Rensvold, 2002; Meade et al., 2008).

Of interesting note in general, ΔMNCI did not seem to be affected by sample size, factor loading magnitude, and number of factor's indicators, resulting in relatively similar 5th and 1st percentiles of ΔMNCI across all conditions. As shown in Figure 5, the 1st percentiles of ΔMNCI seemed to be very similar across different levels of sample size, factor loading magnitude, and number of factor's indicators. As a result, the ΔMNCI equivalent of Figure 4 is not reported because it essentially features nearly overlapping horizontal lines.

Structural Noninvariance Study

Misfit Condition 1 (Large Discrepancy). Regardless of the condition for the factor loading magnitude, ΔMNCI always surpassed −0.02 in Condition 1, indicating that with a large discrepancy between groups, ΔMNCI was able to consistently detect the structural noninvariance for the sample size used in this study. However, ΔCFI did not perform as well. As shown in Figure 6 (the dashed black line is a reference for the Cheung and Rensvold recommendation for ΔMNCI; the dashed gray line is a reference for ΔCFI), not until the factor loading magnitudes were about 0.55 was ΔCFI able to detect that constraining the structural parameters between groups was a misspecification when using criteria set forth by Cheung and Rensvold (2002). Similar to findings from Hancock and Mueller (2011) presented Figure 1, ΔCFI had a slight quadratic pattern for higher measurement quality conditions. As the factor loading magnitude increased, the indices became more and more sensitive to the misfit as seen by the upward trend for both values. The line representing ΔMNCI is not shown for larger factor loading magnitude values in Figure 6 because the values were too large to display simultaneously with ΔCFI.

Misfit Condition 2 (Medium Discrepancy). ΔMNCI exceeded the −0.02 suggested cutoff when the standardized factor loadings were 0.55 or greater, correctly detecting that structural parameter constraints equal did not fit well. On the other hand, ΔCFI was not able to detect the misfit in the constrained structural parameters based on criteria set forth in Cheung and Rensvold (2002) until factor loadings became quite

Table 3. Mean, Standard Deviation, 5th Percentile, and 1st Percentile of ΔCFI and ΔMNCI (Number of Indicators per Factor = 3).

		ΔCFI				ΔMNCI			
N	Loading	Mean	*SD*	5th	1st	Mean	*SD*	5th	1st
$n_1 = n_2 = 100$	.40	−.0126	.1401	−.1650	−.3184	.0011	.0040	−.0060	−.0128
	.45	−.0241	.0667	−.1462	−.2768	.0006	.0045	−.0075	−.0181
	.50	−.0200	.0433	−.1066	−.2068	.0003	.0047	−.0087	−.0189
	.55	−.0155	.0321	−.0844	−.1590	.0000	.0050	−.0100	−.0199
	.60	−.0115	.0237	−.0638	−.1177	−.0001	.0052	−.0107	−.0206
	.65	−.0083	.0172	−.0457	−.0818	−.0001	.0052	−.0111	−.0199
	.70	−.0061	.0126	−.0333	−.0579	−.0002	.0052	−.0111	−.0195
	.75	−.0045	.0093	−.0253	−.0446	−.0002	.0052	−.0111	−.0190
	.80	−.0033	.0068	−.0186	−.0324	−.0002	.0052	−.0115	−.0191
	.85	−.0024	.0050	−.0128	−.0241	−.0002	.0052	−.0109	−.0191
	.90	−.0017	.0035	−.0091	−.0168	−.0002	.0052	−.0107	−.0191
	.95	−.0011	.0023	−.0056	−.0114	−.0002	.0052	−.0104	−.0194
$n_1 = n_2 = 200$	.40	−.0207	.0468	−.1133	−.2078	.0005	.0043	−.0080	−.0143
	.45	−.0156	.0339	−.0823	−.1633	.0002	.0046	−.0091	−.0155
	.50	−.0109	.0227	−.0581	−.1101	.0000	.0048	−.0096	−.0162
	.55	−.0075	.0154	−.0401	−.0745	.0000	.0048	−.0096	−.0161
	.60	−.0053	.0108	−.0288	−.0482	.0000	.0048	−.0093	−.0159
	.65	−.0038	.0078	−.0201	−.0347	.0000	.0048	−.0095	−.0157
	.70	−.0028	.0057	−.0148	−.0265	.0000	.0048	−.0099	−.0176
	.75	−.0021	.0043	−.0106	−.0206	−.0001	.0049	−.0099	−.0183
	.80	−.0016	.0032	−.0080	−.0153	−.0001	.0050	−.0098	−.0193
	.85	−.0012	.0024	−.0060	−.0114	−.0001	.0051	−.0101	−.0196
	.90	−.0008	.0017	−.0044	−.0081	−.0002	.0051	−.0101	−.0190
	.95	−.0005	.0011	−.0028	−.0054	−.0002	.0052	−.0102	−.0192
$n_1 = n_2 = 300$	.40	−.0156	.0318	−.0913	−.1432	.0000	.0015	−.0033	−.0055
	.45	−.0108	.0222	−.0608	−.0979	.0000	.0016	−.0035	−.0061
	.50	−.0073	.0150	−.0418	−.0656	.0000	.0017	−.0036	−.0059
	.55	−.0050	.0104	−.0288	−.0455	.0000	.0017	−.0036	−.0059
	.60	−.0036	.0074	−.0204	−.0327	.0000	.0017	−.0036	−.0059
	.65	−.0026	.0054	−.0148	−.0232	.0000	.0017	−.0036	−.0059
	.70	−.0019	.0040	−.0111	−.0173	.0000	.0017	−.0037	−.0058
	.75	−.0014	.0030	−.0084	−.0125	.0000	.0017	−.0037	−.0058
	.80	−.0011	.0022	−.0062	−.0095	.0000	.0017	−.0037	−.0058
	.85	−.0008	.0017	−.0044	−.0072	−.0001	.0017	−.0037	−.0059
	.90	−.0006	.0012	−.0033	−.0052	−.0001	.0018	−.0039	−.0061
	.95	−.0004	.0008	−.0022	−.0036	−.0001	.0018	−.0040	−.0064
$n_1 = n_2 = 600$	.40	−.0090	.0183	−.0534	−.0795	.0000	.0016	−.0038	−.0050
	.45	−.0055	.0117	−.0321	−.0530	.0000	.0008	−.0018	−.0029
	.50	−.0037	.0078	−.0216	−.0347	.0000	.0008	−.0018	−.0031
	.55	−.0026	.0054	−.0145	−.0242	.0000	.0009	−.0018	−.0032
	.60	−.0019	.0038	−.0100	−.0171	.0000	.0009	−.0018	−.0031
	.65	−.0014	.0028	−.0075	−.0123	.0000	.0009	−.0018	−.0031
	.70	−.0010	.0021	−.0055	−.0091	.0000	.0009	−.0018	−.0033
	.75	−.0008	.0015	−.0042	−.0068	.0000	.0009	−.0018	−.0033
	.80	−.0006	.0012	−.0031	−.0055	.0000	.0009	−.0019	−.0034
	.85	−.0004	.0009	−.0023	−.0042	.0000	.0009	−.0019	−.0036
	.90	−.0003	.0006	−.0016	−.0029	−.0001	.0009	−.0019	−.0036
	.95	−.0002	.0004	−.0011	−.0019	−.0001	.0009	−.0020	−.0035

(continued)

Table 3. (continued)

		ΔCFI				ΔMNCI			
N	Loading	Mean	*SD*	5th	1st	Mean	*SD*	5th	1st
$n_1 = n_2 = 1,000$	.40	−.0051	.0107	−.0281	−.0499	.0000	.0005	−.0009	−.0020
	.45	−.0032	.0069	−.0174	−.0317	.0000	.0005	−.0010	−.0019
	.50	−.0022	.0046	−.0115	−.0211	.0000	.0005	−.0010	−.0019
	.55	−.0015	.0032	−.0080	−.0146	.0000	.0005	−.0010	−.0019
	.60	−.0011	.0023	−.0058	−.0106	.0000	.0005	−.0010	−.0019
	.65	−.0008	.0017	−.0042	−.0078	.0000	.0005	−.0011	−.0020
	.70	−.0006	.0012	−.0031	−.0058	.0000	.0005	−.0011	−.0020
	.75	−.0004	.0009	−.0024	−.0044	.0000	.0005	−.0011	−.0020
	.80	−.0003	.0007	−.0018	−.0033	.0000	.0005	−.0011	−.0020
	.85	−.0002	.0005	−.0013	−.0024	.0000	.0005	−.0011	−.0020
	.90	−.0002	.0004	−.0010	−.0017	.0000	.0005	−.0011	−.0020
	.95	−.0001	.0002	−.0006	−.0011	.0000	.0005	−.0012	−.0019

Table 4. Mean, Standard Deviation, 5th Percentile, and 1st Percentile of ΔCFI and ΔMNCI (Number of Indicators per Factor = 5).

		ΔCFI				ΔMNCI			
N	Loading	Mean	*SD*	5th	1st	Mean	*SD*	5th	1st
$n_1 = n_2 = 100$	.40	−.0110	.0505	−.1074	−.2213	−.0005	.0075	−.0152	−.0245
	.45	−.0061	.0341	−.0655	−.1412	−.0003	.0074	−.0143	−.0232
	.50	−.0035	.0234	−.0440	−.0925	.0000	.0073	−.0143	−.0225
	.55	−.0022	.0165	−.0327	−.0633	.0001	.0072	−.0144	−.0227
	.60	−.0014	.0120	−.0241	−.0454	.0002	.0071	−.0130	−.0220
	.65	−.0010	.0089	−.0174	−.0318	.0003	.0071	−.0133	−.0228
	.70	−.0006	.0066	−.0133	−.0246	.0003	.007	−.0136	−.0229
	.75	−.0005	.0050	−.0099	−.0189	.0004	.0069	−.0130	−.0220
	.80	−.0003	.0037	−.0074	−.0139	.0004	.0069	−.0132	−.0212
	.85	−.0002	.0027	−.0055	−.0103	.0004	.0068	−.0128	−.0203
	.90	−.0002	.0019	−.0038	−.0074	.0004	.0067	−.0131	−.0202
	.95	−.0001	.0013	−.0026	−.0049	.0004	.0066	−.0127	−.0202
$n_1 = n_2 = 200$	.40	−.0045	.0249	−.0532	−.1007	−.0002	.0036	−.1007	−.0108
	.45	−.0027	.0162	−.0343	−.0666	−.0001	.0036	−.0666	−.0114
	.50	−.0018	.0110	−.0235	−.0445	.0000	.0035	−.0445	−.0118
	.55	−.0012	.0078	−.0163	−.0316	.0000	.0035	−.0316	−.0118
	.60	−.0009	.0057	−.0118	−.0218	.0000	.0035	−.0218	−.0114
	.65	−.0007	.0042	−.0091	−.0165	.0000	.0035	−.0165	−.0116
	.70	−.0005	.0032	−.0069	−.0122	.0000	.0035	−.0122	−.0116
	.75	−.0004	.0024	−.0054	−.0091	.0000	.0035	−.0091	−.0123
	.80	−.0003	.0018	−.0041	−.0071	.0000	.0035	−.0071	−.0119
	.85	−.0002	.0014	−.0030	−.0056	−.0001	.0035	−.0056	−.0115
	.90	−.0002	.0010	−.0021	−.0041	−.0001	.0035	−.0041	−.0120
	.95	−.0001	.0007	−.0015	−.0027	−.0002	.0036	−.0027	−.0119

(continued)

Table 4. (continued)

		ΔCFI				ΔMNCI			
N	Loading	Mean	*SD*	5th	1st	Mean	*SD*	5th	1st
$n_1 = n_2 = 300$	.40	−.0038	.0190	−.0395	−.0790	−.0001	.0025	−.0050	−.0083
	.45	−.0023	.0123	−.0264	−.0509	−.0001	.0025	−.0049	−.0080
	.50	−.0015	.0084	−.0182	−.0345	−.0001	.0025	−.0051	−.0079
	.55	−.0011	.0059	−.0126	−.0242	−.0001	.0025	−.0052	−.0082
	.60	−.0008	.0043	−.0094	−.0175	−.0001	.0025	−.0052	−.0082
	.65	−.0006	.0031	−.0070	−.0132	−.0001	.0025	−.0050	−.0083
	.70	−.0004	.0024	−.0055	−.0104	−.0001	.0025	−.0052	−.0084
	.75	−.0003	.0018	−.0042	−.0080	−.0001	.0025	−.0048	−.0087
	.80	−.0003	.0013	−.0030	−.0060	−.0001	.0025	−.0050	−.0089
	.85	−.0002	.0010	−.0022	−.0046	−.0001	.0025	−.0051	−.0091
	.90	−.0002	.0007	−.0016	−.0033	−.0001	.0025	−.0054	−.0091
	.95	−.0001	.0005	−.0011	−.0023	−.0002	.0026	−.0056	−.0092
$n_1 = n_2 = 600$	.40	−.0014	.0092	−.0200	−.0406	.0000	.0013	−.0024	−.0044
	.45	−.0009	.0060	−.0125	−.0261	.0000	.0012	−.0025	−.0043
	.50	−.0006	.0041	−.0085	−.0180	.0000	.0012	−.0025	−.0042
	.55	−.0004	.0029	−.0058	−.0128	.0000	.0012	−.0025	−.0043
	.60	−.0003	.0021	−.0044	−.0093	.0000	.0012	−.0025	−.0043
	.65	−.0002	.0015	−.0033	−.0067	.0000	.0012	−.0025	−.0042
	.70	−.0002	.0011	−.0025	−.0049	.0000	.0012	−.0025	−.0041
	.75	−.0001	.0009	−.0018	−.0037	.0000	.0012	−.0024	−.0041
	.80	−.0001	.0006	−.0013	−.0029	.0000	.0012	−.0024	−.0043
	.85	−.0001	.0005	−.0010	−.0021	.0000	.0012	−.0023	−.0041
	.90	−.0001	.0004	−.0007	−.0016	.0000	.0012	−.0022	−.0042
	.95	−.0000	.0002	−.0005	−.0011	.0000	.0012	−.0023	−.0041
$n_1 = n_2 = 1,000$	.40	−.0011	.0055	−.0118	−.0220	−.0001	.0008	−.0015	−.0026
	.45	−.0007	.0035	−.0077	−.0138	−.0001	.0008	−.0015	−.0025
	.50	−.0005	.0024	−.0053	−.0095	−.0001	.0008	−.0015	−.0024
	.55	−.0003	.0017	−.0038	−.0069	−.0001	.0007	−.0015	−.0024
	.60	−.0002	.0012	−.0027	−.0046	−.0001	.0007	−.0015	−.0023
	.65	−.0002	.0009	−.0021	−.0033	−.0001	.0007	−.0015	−.0022
	.70	−.0001	.0007	−.0016	−.0024	−.0001	.0007	−.0015	−.0021
	.75	−.0001	.0005	−.0012	−.0018	−.0001	.0007	−.0016	−.0022
	.80	−.0001	.0004	−.0010	−.0014	−.0001	.0007	−.0016	−.0022
	.85	−.0001	.0003	−.0007	−.0011	−.0001	.0007	−.0016	−.0022
	.90	.0000	.0002	−.0005	−.0008	−.0001	.0008	−.0016	−.0022
	.95	.0000	.0001	−.0003	−.0006	−.0001	.0008	−.0015	−.0024

high. As seen in Figure 7, ΔCFI required factor loadings of 0.85 before detecting that constraining the structural parameters did not fit the data well.

Misfit Condition 3 (Small Discrepancy). When the discrepancy between the groups was smaller (standardized structural paths of 0.60 vs. 0.80 across groups), ΔMNCI was only able to detect misfit for factor loading magnitudes of 0.75 or above. ΔCFI was not able to detect misfit based on criteria set forth in Cheung and Rensvold (2002)

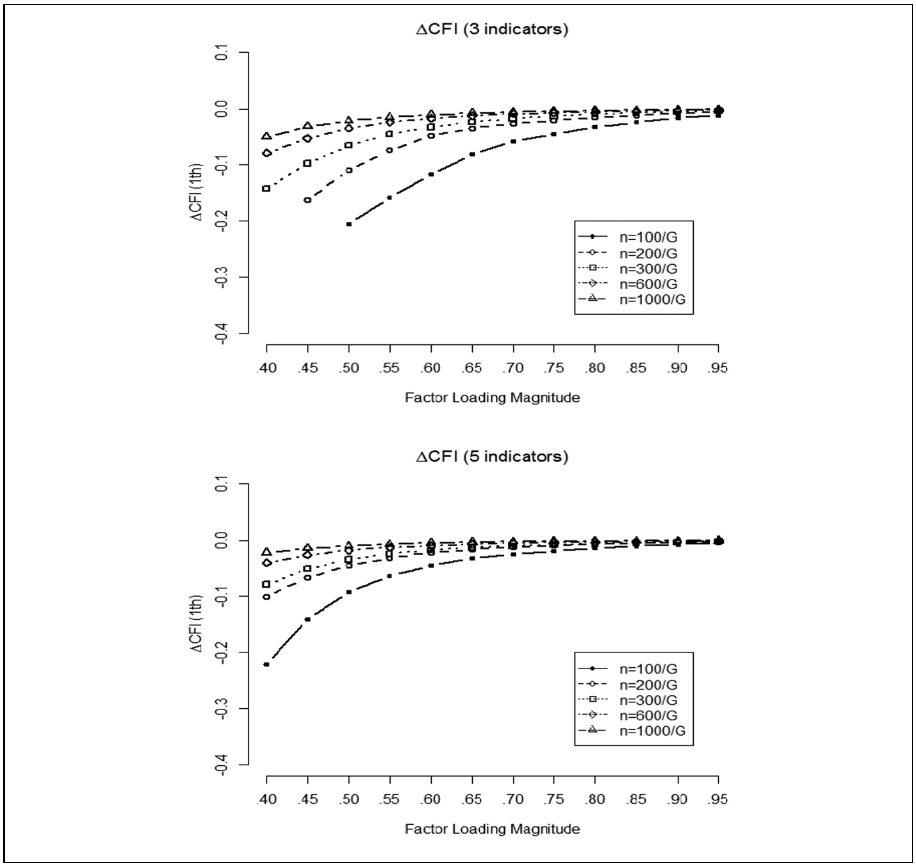

Figure 3. Changes in 1st percentile of ΔCFI.

regardless of the condition for the factor loading magnitude. Figure 8 shows the ΔGOFs across conditions for Condition 3.

Asymptotic Standard Errors and Test Statistics. In addition to implications for assessing data-model fit over the range of factor loading magnitude, the standard error estimates of the structural parameters also are affected by changes in the magnitude of the standardized factor loadings. More specifically, the standard error estimates are much larger when the magnitude of the loadings is smaller, which results in much smaller *Z*-values, which could affect inferences made from the model. Table 5 shows the estimate from the constrained structural estimate from Factor 1 to Factor 4, its standard error, and the *Z*-value for the 0.40 and 0.95 factor loading conditions. The structural parameters from Factor 2 to Factor 5 and Factor 3 to Factor 6 were quite similar, so they are not reported for brevity. Although the binary null hypothesis decision would be congruent in either case in this example, the difference in the *Z*-values

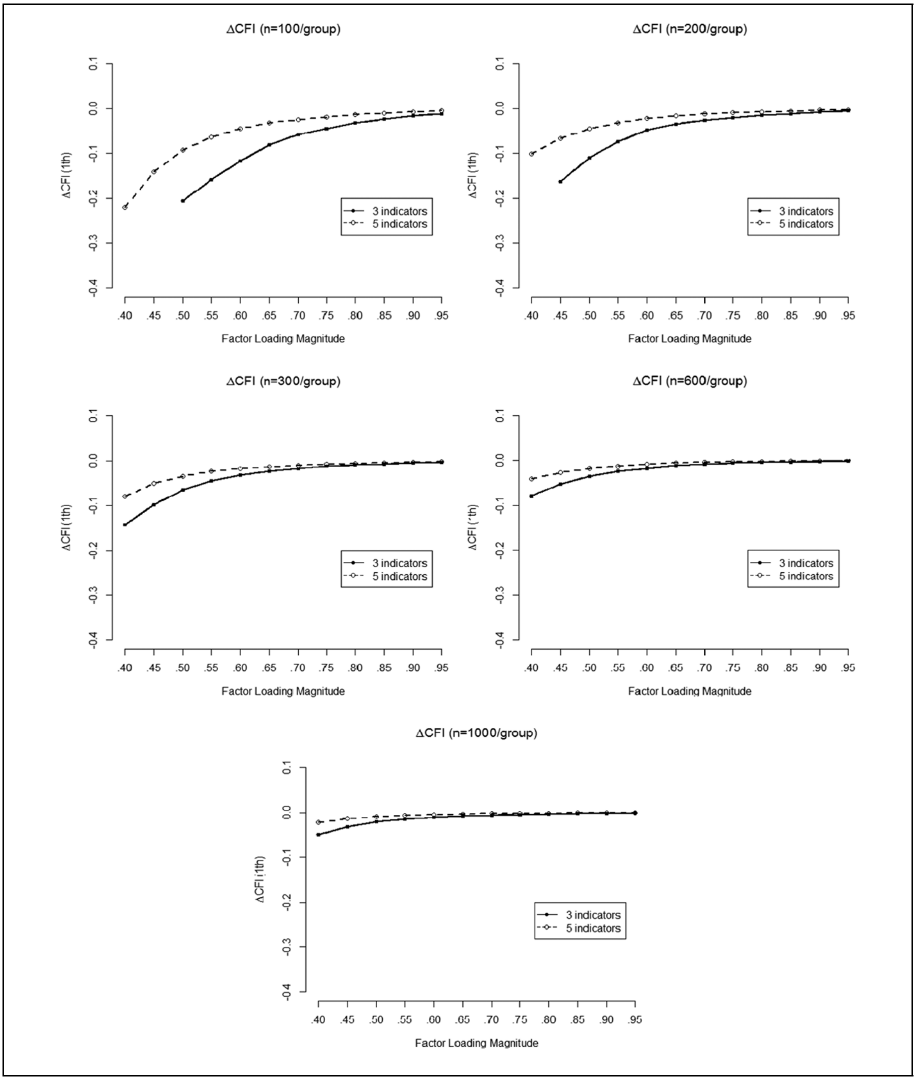

Figure 4. Changes in 1st percentile of ΔCFI across number of factor's indicator.

is by no means trivial and models using smaller sample sizes could easily results in incongruent decisions solely based on measurement quality.

Conclusions, Discussion, and Recommendations

The main purpose of this study was to investigate the role of factor loading magnitude, number of factor indicators, and sample size on recommended cutoff values for

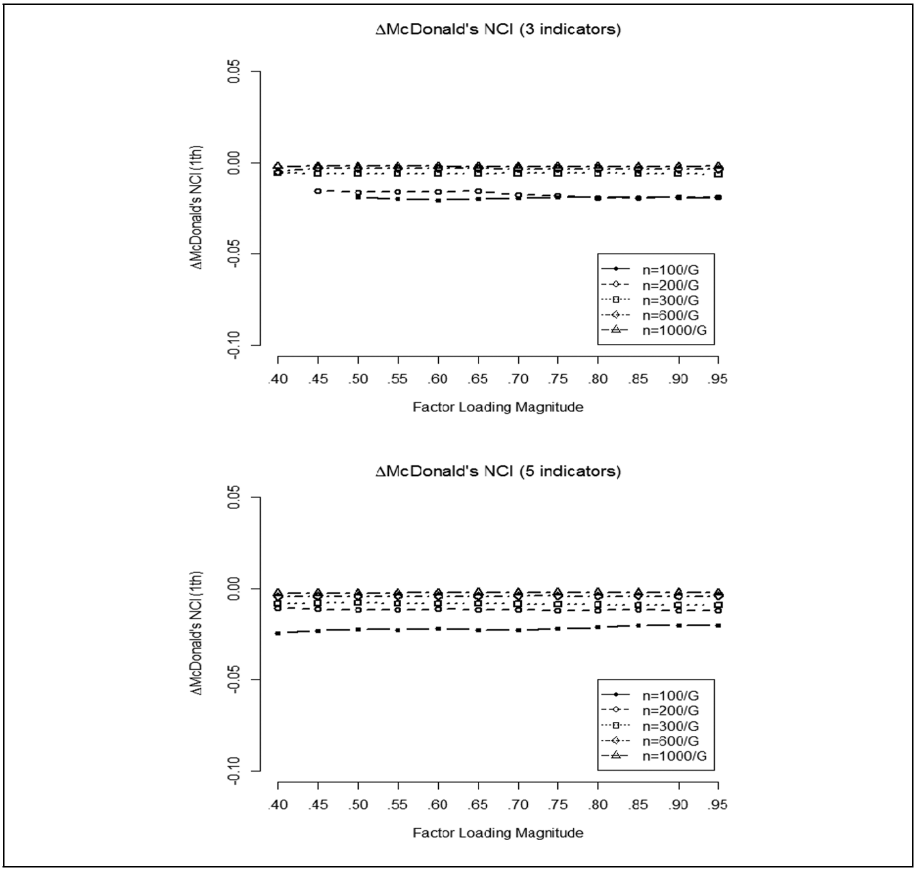

Figure 5. Changes in 1st percentile of ΔMNCI.

assessing measurement invariance with ΔCFI and ΔMNCI, with a secondary focus on the ability of ΔCFI and ΔMNCI to detect structural noninvariance across different levels of measurement quality. Most notably and most unexpectedly based on Hancock and Mueller (2011), ΔMNCI was found to be essentially unaffected by changes in measurement quality (and sample size) when testing measurement invariance, indicating that a single fit index criteria may be relatively stable across a wide range of conditions. Although there is debate within the methodological about the utility of using cutoff values to determine data-model fit (see, e.g., Barrett, 2007; Hayduk, Cummings, Boadu, Pazderka-Robinson, & Boulianne, 2007), should researchers subscribe to this philosophy (with all appropriate precautions), the recommended empirically derived cutoff values across conditions of measurement quality for ΔMNCI are −0.007 and −0.01 for the 5th and 1st percentile, respectively, based on the results of the simulation performed here.

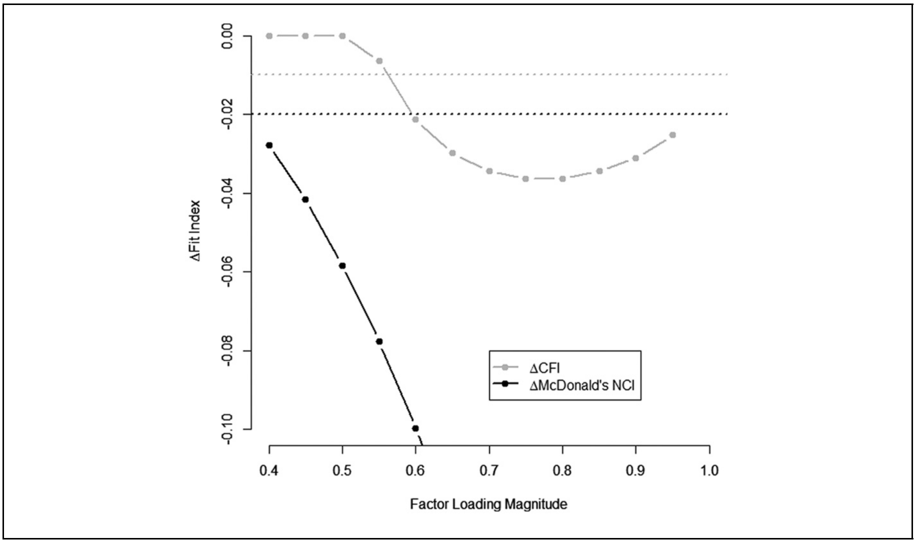

Figure 6. ΔGOF values by standardized factor loading, Condition 1.

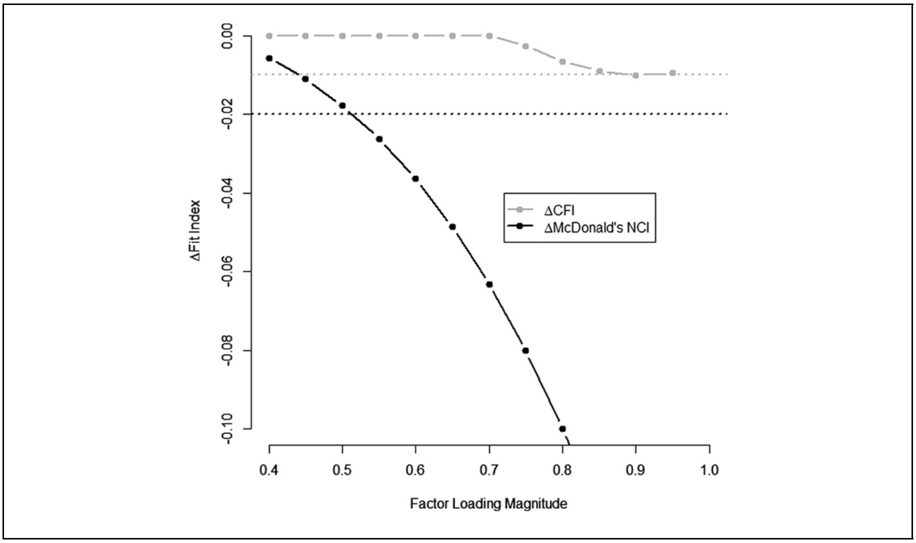

Figure 7. ΔGOF values by standardized factor loading, Condition 2.

This study also found little effect of the number of indicators per factor and sample size on ΔMNCI. Inconsistent with Meade et al. (2008), the number of indicators per factor was found to have little impact on ΔMNCI with an ω^2 effect size less than

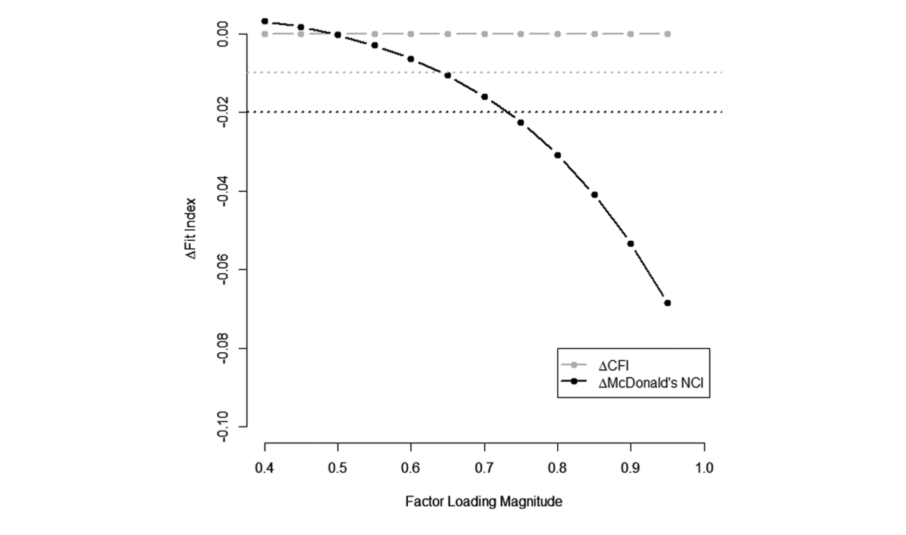

Figure 8. ΔGOF values by standardized factor loading, Condition 3.

Table 5. Structural Parameter Estimate, Standard Error, and *Z*-Values for Two Different Loading Conditions Across Misfit Conditions.

	Loading = 0.40			Loading = 0.95		
Misfit condition	Estimate	*SE*	*Z*-value	Estimate	*SE*	*Z*-Value
1	0.45	0.067	6.98	0.45	0.021	21.18
2	0.60	0.090	6.72	0.60	0.020	30.56
3	0.70	0.096	7.33	0.70	0.018	38.46

<0.001 in this study (Meade et al. reported an ω^2 effect size of 0.063 for number of indicators per factor). With regard to structural invariance, ΔMNCI performed best in the population analysis, although, unlike tests for measurement invariance, there was a noticeable impact of measurement quality on ΔMNCI and the magnitude of values steadily increased as measurement quality increased. However, ΔMNCI was able to detect structural noninvariance far better than ΔCFI (especially when the difference between groups is rather small) and ΔMNCI is thus recommended when testing either measurement or structural invariance with the caveat that a single fit index value cannot apply broadly across factor loading conditions when testing structural invariance. A more comprehensive study would be needed to more definitively support use of ΔMNCI for structural invariance and what values of ΔMNCI are indicative of invariance or noninvariance in the structural portion of the model.

Inconsistent with the findings of previous studies (Cheung & Rensvold, 2002; Meade et al., 2008), this study found that ΔCFI appeared to be affected by sample size. This effect may not have been detected in previous studies because ΔCFI tended to be more affected by sample size when models had low measurement quality (as noted by the significant interaction effect in the ANOVA) and previous studies employed models with consistently high measurement quality. When sample size was small, means of ΔCFI had relatively larger magnitudes and the variability of ΔCFI was relatively larger, resulting in relatively larger magnitudes of the 5th and 1st percentiles (i.e., cutoff values) of ΔCFI. When sample size was large, however, means of ΔCFI were relatively smaller in magnitude and the variability of ΔCFI became relatively smaller, yielding cutoff values of ΔCFI that were smaller in magnitude (closer to 0). It was found that increases in factor loading magnitude generally led to means with smaller magnitude and smaller standard deviation values of ΔCFI across all sample sizes, resulting in cutoff values for ΔCFI with smaller magnitudes. Thus, under a broader range of conditions (particularly for measurement quality), ΔCFI does not appear to follow the fourth desirable property of ΔDOFs advanced by Cheung and Rensvold (2002). Using ΔCFI to assess invariance (either measurement or structural) is not recommended because, as demonstrated in the simulation, values suggestive of invariance change markedly as a function of factor loading magnitude, the number of indicators per factor, and the sample size: the differential values of ΔCFI across measurement quality conditions make a single cutoff difficult to pinpoint because values obtained and suggested in methodological studies may not generalize well to other contexts.

As limitations, only a few ΔGOFs were investigated and other model fit indices such as the Jöreskog-Sörbom GFI (Jöreskog & Sörbom, 1993), adjusted GFI (Jöreskog & Sörbom, 1993), and standardized root mean squared residual (Jöreskog & Sörbom, 1993) have not been investigated. ΔGOFs were chosen based those that previous studies found to have desirable properties, but future studies could examine how measurement quality relates to changes in additional model fit indices in invariance testing under a broader set of conditions in the event that additional indices may perform well. Another limitation is that this study examined the Type I error rate for measurement invariance and Type II errors for structural invariance (not Type II error *rate* because a population analysis was used). It is recommended, thus, that future studies investigate power of ΔGOFs under various degrees of measurement and structural noninvariance conditions as well (although only ΔMNCI has demonstrated reasonable performance across a variety of conditions, so power may be a moot point if only one GOF exhibits desirable properties). The simulation study and population analysis also did not include trivial misspecifications when data were generated. Analysis models used in the studies had perfect fit, which is not entirely reminiscent of models seen in practice. Therefore, the results might be considered a baseline for the effect of sample size and measurement quality on ΔGOFs. This study was intended more so to familiarize readers with the associated issues with ΔGOFs and varying measurement quality in multiple group models because this has yet to receive attention in the methodological literature; future research to expand on this study (e.g., different models, more indices, including trivial misspecifications) would be valuable given the widespread interest in invariance testing that exists across many disciplines.

Overall, the effect Hancock and Mueller (2011) found in single group models whereby higher measurement quality leads to seemingly worse data-model fit was found to generalize to multiple group scenarios, as expected, with the exception of ΔMNCI for testing invariance in the measurement model. However, seeing as measurement invariance testing is the most widely implemented type of invariance test, this finding is a potentially valuable one for applied researchers. Methodological studies that provide the current conventional cutoffs for poor or acceptable data-model fit may not be as widely applicable as applied researchers presume (although cutoffs were admittedly not intended to be appropriate in all circumstances; see Hu & Bentler, 1998, 1999) and researchers should heed quality of measurement when testing invariance of both the measurement and structural model across groups. If this information is not taken into account, in the specific context of multiple group analyses, the failure to detect that parameters should not be constrained can have a detrimental impact on inferences in practical applied scenarios. Constructs might be erroneously considered to function similarly between different countries, cultures, or demographic groups not as a result of more theoretical or modeling choices but simply as a result of poor measurement quality, which not only would result in flawed inferences from the model such as tests, assessments, or instruments being inappropriately administered as interchangeable between groups, but would also quell further research efforts into explaining possible differences between groups or revising instruments to so that they are more broadly administrable.

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) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Some sources (e.g., van de Schoot, Lugtig, & Hox, 2012) also mention a fourth, more basic type called configural invariance, which assumes that the same theoretical model holds across all relevant groups. Under the four-type classification, configural invariance is prerequisite to weak, strong, and strict invariance.
2. Comparisons of mean structures (if present) may not be warranted with only weak invariance. If one can only establish weak invariance, then latent means may be due to

differences in manifest intercepts, which are not constrained with weak invariance (Brown, 2006).

3. The cutoff values are negative because higher values of these indices are indicative of better fit. ΔGOFs are calculated by subtracting the unconstrained GOF value (which will have equal or greater fit and will therefore be higher) from the constrained GOF (which will have equal or worse fit and will therefore be lower), resulting in negative values.
4. ω^2 values between 0.01 and 0.06 are typically interpreted as being ‘‘small’’ effects (Cohen, 1988).

References

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, *42*, 815-824.

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

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema*, *20*, 872-882.

Byrne, B. M. (2013). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Hillsdale, NJ: Routledge.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*, 456-466.

Campbell-Sills, L., Liverant, G. I., & Brown, T. A. (2004). Psychometric evaluation of the behavioral inhibition/behavioral activation scales in a large sample of outpatients with anxiety and mood disorders. *Psychological Assessment*, *16*, 244-254.

Chen, H., Keith, T. Z., Weiss, L., Zhu, J., & Li, Y. (2010). Testing for multigroup invariance of second-order WISC-IV structure across China, Hong Kong, Macau, and Taiwan. *Personality and Individual Differences*, *49*, 677-682.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*, 464-504.

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, *25*, 1-27.

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology*, *31*, 187-212.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Routledge.

Finch, W. H., & French, B. F. (2013). *Model invariance testing under different level of invariance*. Paper presented at the Modern Modeling Methods Conference, Storrs, CT.

Fonseca-Pedrero, E., Wells, C., Paino, M., Lemos-Giráldez, S., Villazón-García, Ú., Sierra, S., & Muñiz, J. (2010). Measurement invariance of the Reynolds Depression Adolescent Scale across gender and age. *International Journal of Testing*, *10*, 133-148.

Fornell, C, & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*, 39-50.

Gagné, P. E., & Hancock, G. R. (2006). Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research*, *41*, 65-83.

Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika*, *66*, 373-388.

Hancock, G. R., & French, B. F. (2013). Power analysis in covariance structure models. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 117-159). Charlotte, NC: Information Age.

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future—A festschrift in honor of Karl Joreskog*. Lincolnwood, IL: Scientific Software International.

Hancock, G. R., & Mueller, R. O. (2011). The reliability paradox in assessing structural relations within covariance structure models. *Educational and Psychological Measurement*, *71*, 306-324.

Hayduk, L., Cummings, 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.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and meta-analysis. *Sociological Methods & Research*, *26*, 329-367.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*, 117-144.

Horn, J. L., McArdle, J., & Mason, R. (1983). When is invariance not invariant: A practical scientist's look at the ethereal concept of factor invariance. *The Southern Psychologist*, *1*, 179-188.

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*, 424-453.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1-55.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*, 409-426.

Jöreskog, K., & Sörbom, D. (1993). *LISREL 8: User's reference guide*. Chicago, IL: Scientific Software International.

Karcher, M. J., & Sass, D. (2010). A multicultural assessment of adolescent connectedness: testing measurement invariance across gender and ethnicity. *Journal of Counseling Psychology*, *57*, 274-289.

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

Little, T. D., Card, N. A., Slegers, D. W., & Ledford, E. C. (2007). Representing contextual effects in multiple-group MACS models. In T. D. Little, J. A. Bovaird, & N. A. Card (Eds.), *Modeling contextual effects in longitudinal studies* (pp. 121-147). Mahwah, NJ: Erlbaum.

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

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.

Meredith, W. (1993). Measurement invariance, factor analysis and factor invariance. *Psychometrika*, *58*, 525-544.

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, *44*, S69-S77.

Millsap, R. E., & Meredith, W. (2004, May). *Factorial invariance: Historical trends and new developments*. Paper presented at the Factor Analysis at 100 Conference, Chapel Hill, NC.

Moura, O., dos Santos, R. A., Rocha, M., & Matos, P. M. (2010). Children's Perception of Interparental Conflict Scale (CPIC): Factor structure and invariance across adolescents and emerging adults. *International Journal of Testing*, *10*, 364-382.

Muthén, L., & Muthén, B. (2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Saris, W. E., Satorra, A., & Van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling*, *16*, 561-582.

Savickas, M. L., & Porfeli, E. J. (2012). Career Adapt-Abilities Scale: Construction, reliability, and measurement equivalence across 13 countries. *Journal of Vocational Behavior*, *80*, 661-673.

Teo, T., Lee, C. B., Chai, C. S., & Wong, S. L. (2009). Assessing the intention to use technology among pre-service teachers in Singapore and Malaysia: A multigroup invariance analysis of the technology acceptance model (TAM). *Computers & Education*, *53*, 1000-1009.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4-7.

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*, 486-492.

van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, *54*, 119-135.

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research and Evaluation*, *12*, 1-26.

Zhang, B., Fokkema, M., Cuijpers, P., Li, J., Smits, N., & Beekman, A. (2011). Measurement invariance of the Center for Epidemiological Studies–Depression Scale (CES-D) among Chinese and Dutch elderly. *BMC Medical Research Methodology*, *11*, 74.