

Equivalence Testing for Factor Invariance Assessment with Categorical Indicators



W. Holmes Finch and Brian F. French

Abstract Factorial invariance assessment is central in the development of educational and psychological instruments. Establishing factor structure invariance is key for building a strong validity argument, and establishing the fairness of score use. Fit indices and guidelines for judging a lack of invariance is an ever-developing line of research. An equivalence testing approach to invariance assessment, based on the RMSEA has been introduced. Simulation work demonstrated that this technique is effective for identifying loading and intercept noninvariance under a variety of conditions, when indicator variables are continuous and normally distributed. However, in many applications indicators are categorical (e.g., ordinal items). Equivalence testing based on the RMSEA must be adjusted to account for the presence of ordinal data to ensure accuracy of the procedures. The purpose of this simulation study is to investigate the performance of three alternatives for making such adjustments, based on work by Yuan and Bentler (*Sociological Methodology*, 30(1):165–200, 2000) and Maydeu-Olivares and Joe (*Psychometrika* 71(4):713–732, 2006). Equivalence testing procedures based on RMSEA using this adjustment is investigated, and compared with the Chi-square difference test. Manipulated factors include sample size, magnitude of noninvariance, proportion of noninvariant indicators, model parameter (loading or intercept), and number of indicators, and the outcomes of interest were Type I error and power rates. Results demonstrated that the T_3 statistic (Asparouhov & Muthén, 2010) in conjunction with diagonally weighted least squares estimation yielded the most accurate invariance testing outcome.

Keywords Invariance testing · Equivalence test · Categorical indicator

W. Holmes Finch (✉)
Ball State University, Muncie, IN 47306, USA
e-mail: whfinch@bsu.edu

B. F. French (✉)
Washington State University, Pullman, WA 99164, USA

1 Introduction

Social scientists, policy makers, and others make use of scores from psychological scales to make decisions about persons, and groups of people, for a variety of purposes, including hiring, school matriculation, professional licensure, and determinations regarding the need for special educational and psychological services. Given their importance, there must be strong validity evidence for using scores in these ways (American Education Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). One important aspect of providing such evidence is the determination as to whether the measures provide equivalent information for members of different groups in the population, such as males and females, or members of different economic subgroups (Wu, Li, & Zumbo, 2007). Traditionally, such factor invariance (FI) assessments have been made using a Chi-square difference test with multiple group confirmatory analysis (MGCFAs). However, this approach is very sensitive to sample size, so that it might be statistically significant for very minor differences in group parameter values (Yuan & Chan, 2016). Perhaps more importantly, information about the magnitude of any group differences in latent variable model parameters identified is not available (Yuan & Chan, 2016). Yuan and Chan described an alternative approach to FI assessment that is based on equivalence testing. When indicator variables are normally distributed, this equivalence testing based method is an effective tool, yielding accurate results with respect to the invariance (or noninvariance) of the latent variable model (Finch & French, 2018). The purpose of the current simulation study was to extend this earlier work by investigating how the equivalence testing technique performed when the observed indicators were ordinal variables (such as items on a scale), rather than being normally distributed.

1.1 MGCFAs and FI Assessment

FI assessment (Millsap, 2011) refers to a set of nested models with differing levels of cross group equality assumed about the parameters in a latent variable model linking observed indicators (x) to latent variables (ξ). The weakest type of FI is configural invariance (CI), where only the general latent structure (i.e., number of latent variables and correspondence of observed indicators to latent variables) is the same across groups. The next level of FI is measurement invariance (MI), where the factor loading matrix (Λ) is assumed to be equivalent across groups (Kline, 2016; Wicherts & Dolan, 2010). If MI holds, researchers might next assess the equality of the factor model intercepts (τ) across groups (Steenkamp & Baumgartner, 1998), and/or group equality of the unique variances (δ) invariant across groups.

The most common approach for assessing FI is based on the MGCFAs model:

$$x_g = \tau_g + \Lambda_g \xi + \delta_g \tag{1}$$

where

- x_g Observed indicators for group g
- τ_g Threshold parameters for group g
- Λ_g Factor loading matrix for group g
- ξ Latent variable(s)
- δ_g Unique variances of the indicator variables for group g .

The terms in Eq. (1) are as described above, except that the parameters are allowed to vary by group, which is denoted by the g subscript. MGCFA is used to test each type of FI through a series of nested models, which differ in terms of that model parameters that are held equivalent between groups. For example, in order to assess MI, the researcher would constrain factor loadings to be equivalent across groups, thereby replacing Λ_g with Λ in Eq. (1). The fit of the constrained and unconstrained models are then compared using a difference in Chi-square fit statistic value, χ^2_{Δ} . The null hypothesis of this test is that MI is present.

The performance of χ^2_{Δ} for invariance testing has yielded somewhat mixed results. French and Finch (2006) found that for normally distributed indicators and a sample size of no more than 500, χ^2_{Δ} had Type I error rates at the nominal (0.05) level, while also exhibiting relatively high power. Other researchers have reported that χ^2_{Δ} is sensitive to sample size, to a lack of normality in the indicators, and to model misspecification errors, and in such cases may yield inflated Type I error rates when assessing MI (Chen, 2007; Yuan & Chan, 2016; Yuan & Bentler, 2004).

1.2 Factor Invariance with Equivalence Testing

Given these problems associated with using χ^2_{Δ} , Yuan and Chan (2016) proposed an extension of other work designed to assess model fit using an equivalence testing approach (Marcoulides & Yuan, 2017; Yuan, Chan, Marcoulides, & Bentler, 2016) to the assessment of FI using MGCFA. In the case of MGCFA for FI assessment, the null hypothesis is:

$$H_{0I} : (F_{bc0} - F_{b0}) > \varepsilon_0 \quad (2)$$

F_{bc0} is the fit function value for a model where group parameters are constrained to be equal, F_{b0} is the fit function value for a model where group latent variable model parameters are allowed to differ between groups, and ε_0 is the maximum acceptable model misspecification. Rejecting H_{0I} leads to the conclusion that any model misspecification due to constraining factor model parameters to be equal across groups does not greatly degrade model fit vis-à-vis the model where these constraints are relaxed. Therefore, rejecting H_{0I} in the MI equivalence testing framework would indicate that when the groups' factor loadings are constrained to be equal, the difference in fit between the loadings constrained and loadings unconstrained models does not exceed an acceptable level of misfit, as expressed by ε_0 .

Yuan and Chan (2016) showed that the value of ε_0 can be obtained as follows:

$$\varepsilon_0 = \frac{df(RMSEA_0)^2}{m} \quad (3)$$

where

df Model degrees of freedom
 m Number of groups
 $RMSEA_0$ Maximum value of $RMSEA$ that can be tolerated.

For FI assessment, Yuan and Chan recommend using this equivalence testing procedure to characterize the relative degree of noninvariance present in the data, as opposed to making strict hypothesis testing based determinations regarding equivalence or not. In this framework, the degree of model parameter invariance present in the data can be characterized using common guidelines (e.g., MacCallum, Browne, & Sugawara, 1996) to describe the model constraining group parameters to be equal. These guidelines for interpreting values of $RMSEA$ suggest the following fit categories: Excellent fit (<0.01), Close fit (0.01–0.05), Fair fit (0.05–0.08), Mediocre fit (0.08–0.10), and Poor fit (0.10+). Thus, an $RMSEA$ of 0.17 for a model constraining factor loadings to be equal among groups would suggest poor fit of the MI model, meaning that model parameters are likely not equivalent between the groups. Yuan and Chan (2016) found that for the purposes of determining the value of ε_0 , these standard cutoffs for interpreting $RMSEA$ may be too stringent, and thus recommended an alternative approach for obtaining adjusted cutoffs based on the data being analyzed. The interested reader is encouraged to review this earlier paper for a description of how these alternatives are obtained. This equivalence testing approach is effective for assessing the fit of a single model, and for invariance assessment (e.g., Finch & French, 2018; Marcoulides & Yuan, 2017; Yuan & Chan, 2016). However, the performance of the equivalence testing approach to invariance assessment when indicators are categorical and not normally distributed has not been investigated.

1.3 Fit Indices for Categorical Indicators

Yuan and Chan (2016) indicated that the equivalence invariance test was designed for use with normally distributed indicators. However, in many contexts in the social sciences researchers work with ordinal observed variables, such as responses to items on a rating scale. In such cases, the equivalence testing approach may not be appropriate, because calculation of the standard full information χ^2 statistic upon which $RMSEA$ is based is problematic (Maydeu-Olivares & Joe, 2006). In the context of categorical indicators, this statistic relies on the full cross-tabulation of the entire set of categorical indicators (full information), leading to the potential for cell sparsity, and resulting problems in its calculation (Maydeu-Olivares & Joe), which in turn biases the $RMSEA$ estimate.

In order to address these problems caused by sparsity, alternative goodness of fit statistics based on limited information approaches have been proposed for use with latent variable modeling in the context of categorical indicator variables. One set of alternatives is based upon a least squares, rather than maximum likelihood, estimation paradigm. For example, the weighted least squares (WLS) fit function takes the form:

$$F(\theta; W) = (\hat{\rho} - \rho(\theta))' W (\hat{\rho} - \rho(\theta)) \tag{4}$$

where

- $\hat{\rho}$ Sample polychoric correlation matrix for the indicator variables
- $\rho(\theta)$ Model implied polychoric correlation matrix
- W Asymptotic covariance matrix of $\hat{\rho}$.

Given that WLS yields biased estimates and has difficulty in converging when samples are small (Muthén, 1993), the diagonally weighted least squares (DWLS) estimator was proposed (Muthén, du Toit, & Spisic, 1997). DWLS reduces the computational burden and yields less biased parameter estimates for smaller sample sizes by using only the diagonal of W as the weight matrix (Flora & Curran, 2004). When W is the identity matrix, (5) is the unweighted least squares (ULS) estimator. For each of these estimators, a moment corrected goodness of fit statistic, T_3 , can be calculated based upon the fit function, and is asymptotically a Chi-square statistic (Asparouhov & Muthén, 2010). T_3 can then be used to calculate RMSEA, which in turn can be used with the invariance equivalence methodology described above.

An alternative limited information goodness of fit statistic for use with categorical indicators was proposed by Maydeu-Olivares and Joe (2006). This statistic is defined as:

$$M_2^* = N \hat{e}_2 \hat{\Omega}_2 \hat{e}_2 \tag{5}$$

where

- \hat{e}_2 Vector of first and second order residual probabilities.
- $\hat{\Omega}_2 = \Xi_2^{-1} - \Xi_2^{-1} \Delta_2 (\Delta_2' \Xi_2^{-1} \Delta_2)^{-1} \Delta_2' \Xi_2^{-1}$
- Ξ_2 Asymptotic covariance matrix of the first and second order sample proportions
- Δ_2 Matrix of derivatives of the first and second order model implied probabilities with respect to the vector of parameter estimates $\hat{\theta}$.
- M_2^* is asymptotically distributed as a Chi-square statistic, and can be used to calculate RMSEA for use with the invariance equivalence testing approach described above.

1.4 Goals of the Current Study

The goal of the current study was to extend earlier work that investigated the performance of the equivalence testing procedure for normally distributed indicators (Finch & French, 2018). The current study extends this research by examining the performance of T_3 for both DWLS and ULS, as well as M_2 in the context of MI when the indicator variables are categorical.

2 Method

A Monte Carlo simulation study (1000 replications) was utilized to address the study goals. Data simulation was completed in Mplus, version 7.11 (Muthén & Muthén, 1998–2016), and data analyses were conducted using R version 3.3.1 (R Development Core Team, 2016). Data were generated using a single factor confirmatory factor analysis model for 2 groups, where the factor, error variances, and factor variances followed the standard normal distribution, with a mean of 0 and variance of 1. Indicator variables were simulated to be ordinal with 5 categories, with the following pattern of thresholds: $-1, -0.5, 0.5, 1$. Factor loadings were set to 1 for all indicators, unless manipulated to induce measurement noninvariance, as described below. All other model parameters were held invariant between the two groups. The referent indicator method was used to identify the factor models. The following factors were manipulated in the study, and were based upon earlier published work in this area (e.g., Finch & French, 2018).

2.1 Sample Size

Given that sample size has been shown to be important in terms of the performance of the equivalence testing approach, and the χ^2_{Δ} test (Chen, 2007; Finch & French, 2018; French & Finch, 2006), it was manipulated in the current study. Total sample sizes were simulated to be 200, 400, 600, 1000, 1500, or 2000, and were designed to reflect small to large samples.

2.2 Number of Indicator Variables

Either 10 or 20 observed factor indicators were simulated, representing a range of values that might be encountered in practice.

2.3 *Number of Noninvariant Indicators and Magnitude of Measurement Noninvariance*

Measurement noninvariance was simulated by creating group differences in the factor loadings for some observed indicators. For the invariance condition, the difference in factor loadings between the groups was 0 (complete invariance). For the noninvariant cases, loadings were simulated to differ by 0.1, 0.2, 0.3, 0.4, or 0.5. The percent of indicators allowed to be noninvariant was 0, 10, 20, or 30%. As an example of how noninvariance was simulated, in the 10 indicators, 10% noninvariant, 0.1 noninvariance magnitude condition, the factor loading for indicator 2 was set to 0.9 in one group, and kept at 1.0 in the other group.

2.4 *Invariance Assessment Approaches*

For each replication within each simulation condition, invariance was tested using the MGCFA χ^2_{Δ} approach, with T_3 for DWLS (T_{DWLS}) and ULS (T_{ULS}), as well as M_2^* . In addition, the equivalence test method based was also used to assess invariance, with the RMSEA values based upon T_{DWLS} , T_{ULS} , and M_2^* , respectively.

2.5 *Study Outcomes*

The outcomes were the Type I error and power rates of the χ^2_{Δ} tests, and the adjusted equivalence test fit category distribution (Excellent, Close, Fair, Mediocre, or Poor). Analysis of variance (ANOVA) was used to identify statistically significant main effects and interactions of the manipulated conditions with respect to the proportion of cases for which the equivalence testing method identified poor fit. In addition, the partial η^2 effect size was also used to identify ANOVA model terms of interest, such that main effects and interactions of the manipulated conditions had to be statistically significant with partial η^2 value of 0.1 or larger, ensuring that effects accounted for at least 10% of the outcome variance to be deemed important.

3 Results

3.1 *Measurement Invariance Is Present*

The interaction of invariance assessment method by sample size was the only statistically significant model term ($F_{10,8} = 10.527$, $p = 0.001$, $\eta^2 = 0.929$) when invariance was present. The Type I error rate for the T_{DWLS} statistic was the only one

Table 1 Type I error rates and proportion of adjusted equivalence test results in excellent or close range, by sample size and method

N	T_{DWLS}	T_{ULS}	M_2^*	Proportion in excellent/close fit range		
				RMSEA T_{DWLS}	RMSEA T_{ULS}	RMSEA M_2^*
200	0.07	0.08	0.11	0.39	0.12	0.03
400	0.06	0.08	0.10	0.41	0.13	0.03
600	0.04	0.08	0.10	0.57	0.21	0.04
1000	0.05	0.08	0.07	0.80	0.45	0.25
1500	0.05	0.07	0.06	0.84	0.55	0.34
2000	0.05	0.07	0.05	0.92	0.69	0.50

that was in the acceptable range (0.025–0.075) as defined by Bradley (1978), across all sample size conditions (Table 1). For the other two statistics, the samples needed to be at least 1000 (M_2^*) or 1500 (T_{ULS}) in order for the Type I error rates to be in this range. Results in Table 1 show that equivalence testing based on T_{DWLS} had the highest rates in the expected excellent/close fit categories, across sample sizes. This proportion increased concomitantly with increases in sample size. Finally, the proportion in the expected excellent/close fit range was below 0.8 for samples of less than 1000 for T_{DWLS} , which was the best performer in this regard.

3.2 Measurement Invariance Is Not Present

When factor loadings were simulated to differ between the groups, ANOVA found the interactions of equivalence test statistic by number of noninvariant indicator variables by magnitude of group loading difference ($F_{16,234} = 14.389, p < 0.001, \eta^2 = 0.496$), and equivalence test statistic by number of loadings by magnitude of group loading difference ($F_{8,234} = 9.104, p < 0.001, \eta^2 = 0.237$), to be statistically significantly related to the performance of the equivalence test procedure. The first set of results to be examined are those for measurement invariance not present, by method, magnitude of group loading difference, and percent of noninvariant loadings. The proportion of replications in each equivalence testing category for this combination of conditions appear in Fig. 1. It is clear from these results that when the magnitude of group loading differences was 0.3 or more, and 20 or 30% of the indicators were noninvariant between groups, virtually all replications were in the poor fit range (expected outcome given simulated lack of invariance) for all of the methods. Under conditions in which the degree of group difference was less pronounced, the invariance tests based on M_2^* and T_{ULS} tended to indicate worse fit more frequently than did those based on T_{DWLS} . This result was strongest when 30% of the indicators were simulated to have different loadings between groups, and the magnitude of these differences was 0.1 or 0.2. Power results for the χ^2_{Δ} tests appear

Table 2 Power rates for detecting measurement noninvariance for the chi-square difference test by the magnitude of group loading difference, number of noninvariant indicators, and test statistic

Magnitude of group loading difference	Number of noninvariant indicators	T_{DWLS}	T_{ULS}	M_2^*
0.1	1	0.07	0.84	0.15
	2	0.11	0.90	0.28
	3	0.15	0.90	0.29
0.2	1	0.21	0.93	0.44
	2	0.45	0.97	0.63
	3	0.61	0.98	0.67
0.3	1	0.53	0.97	0.63
	2	0.76	0.99	0.82
	3	0.85	0.99	0.87
0.4	1	0.76	0.99	0.79
	2	0.91	0.99	0.92
	3	0.97	1.00	0.95
0.5	1	0.89	0.99	0.89
	2	0.94	0.99	0.94
	3	0.99	1.00	0.990

in Table 2, and demonstrate that T_{ULS} had the highest rates of power across conditions, whereas T_{DWLS} exhibited somewhat lower power than did M_2^* , particularly for lower group loading difference magnitudes, and fewer noninvariant indicators. It is important when interpreting these results to recall that the Type I error rates were inflated under many conditions for each of these statistics, particularly M_2^* and T_{ULS} .

The proportion of replications in each equivalence testing category by magnitude of group loading difference and number of indicators when noninvariance was simulated to be present appear in Fig. 2. These results revealed that with a larger group loading difference there was a higher likelihood of mediocre and poor fit, based on the equivalence test. In addition, with more indicators this effect was magnified for each of the statistics. For example, the proportion of cases in the mediocre and poor fit categories was greater for 20 indicators than for 10, across methods studied here. Power results for the χ^2_{Δ} tests by magnitude of group loading difference and number of indicators appear in Table 3, and are aggregated over the number of non-invariant indicators. Power for all three equivalence testing methods was higher when more indicators were present, and that power for T_{ULS} was the highest across conditions, whereas power for T_{DWLS} was the lowest for the smallest magnitudes of group loading difference.

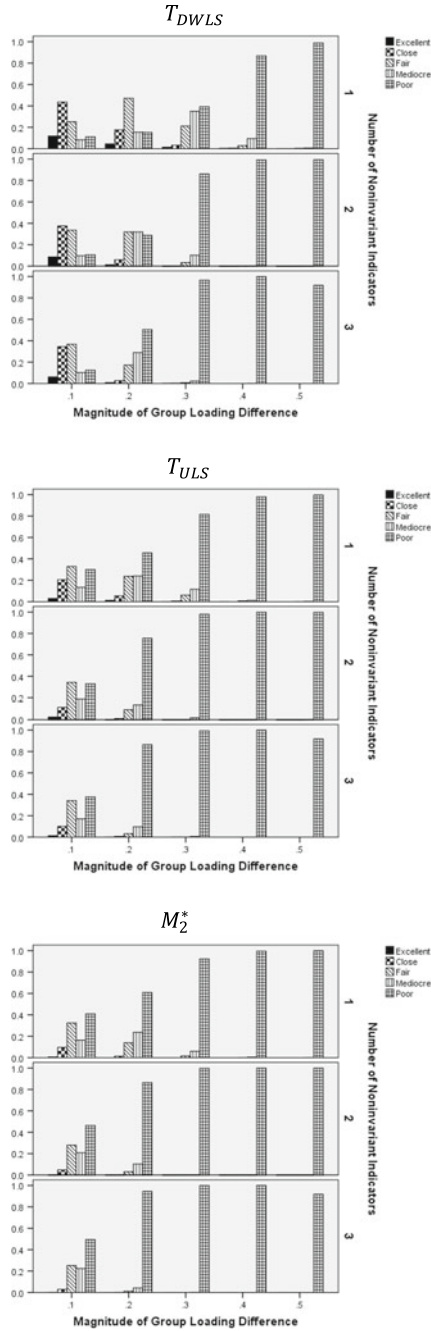


Fig. 1 Proportion of adjusted equivalence test results in each fit category by equivalence statistic, number of noninvariant loadings, and magnitude of group loading difference: noninvariance present

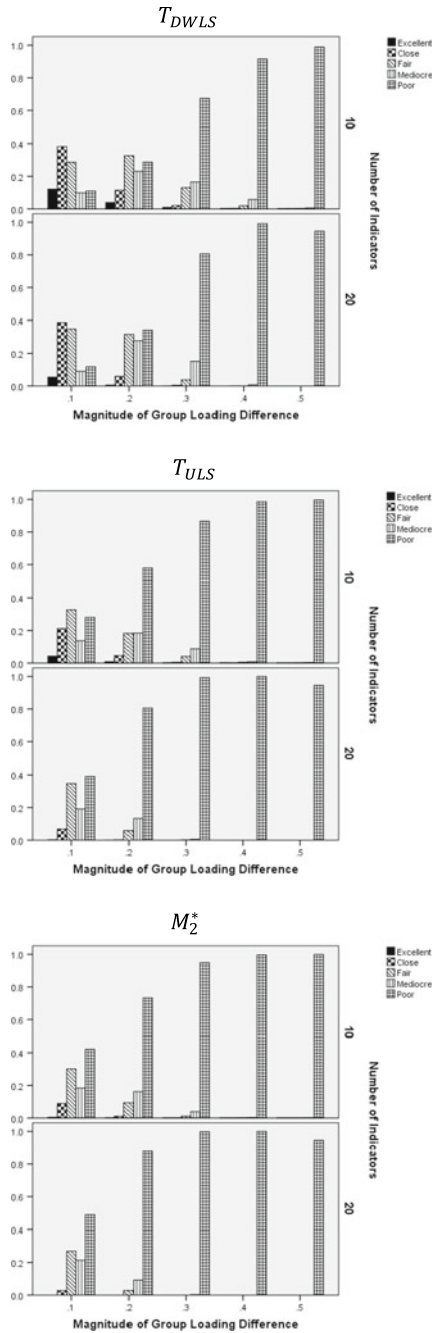


Fig. 2 Proportion of adjusted equivalence test results in each fit category by equivalence statistic, number of indicator variables, and magnitude of group loading difference: noninvariance present

Table 3 Power rates for detecting measurement noninvariance for the chi-square difference test by the magnitude of group loading difference, number of indicators, and test statistic

Magnitude of group loading difference	Number of indicators	T_{DWLS}	T_{ULS}	M_2^*
0.1	10	0.10	0.79	0.30
	20	0.11	0.97	0.39
0.2	10	0.38	0.93	0.58
	20	0.46	0.99	0.68
0.3	10	0.66	0.98	0.74
	20	0.77	0.99	0.81
0.4	10	0.84	0.99	0.85
	20	0.92	1.00	0.92
0.5	10	0.91	0.99	0.91
	20	0.97	1.00	0.97

4 Discussion

The results of this study demonstrated that the equivalence testing procedure based on T_{DWLS} appeared to correctly identify models in which MI held at the highest rates among the methods studied here, while at the same time generally identifying poorly fitting models at a high rate. It is important to note that when the magnitude of group factor loading difference was relatively low (0.2 or less), this statistic was less likely to indicate fair to poor fit than the alternatives studied here. This result could suggest a relative lack of power for this approach, or it could simply reflect the fact that small differences in factor loadings are not indicative of a major lack of equivalence between groups. Finally, the χ^2_{Δ} based approaches exhibited inflated Type I error rates in many cases, and may not be as useful as the equivalence testing approach.

Future research in this area should focus on identifying additional alternatives for calculating RMSEA with categorical indicators. Though T_{DWLS} was the best performer, it was not without problems, particularly for low levels of noninvariance. In addition, future work should include a wider array of indicator categories (e.g., 3, 4, 6, 7), and more complex latent structure (e.g., 2 or 3 factors). Such continued work will allow the invariance literature to continue to expand to address group differences in the measurement of constructs used to make decisions about individuals.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational & Psychological Testing*. Washington, D.C.: American Educational Research Association.
- Asparouhov, T., & Muthén, B. O. (2010). *Simple second order chi-square correction*. Retrieved from: http://www.statmodel.com/download/WLSMV_new_chi21.pdf.
- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 321–339.
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504.
- Finch, W. H., & French, B. F. (2018). A simulation investigation of the performance of invariance assessment using equivalence testing procedures. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(5), 673–686.
- Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466–491.
- French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*, 378–402.
- Kline, R. B. (2016). *Principles and practice of structural equation modeling*. New York: The Guilford Press.
- MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*, 130–149.
- Marcoulides, K. M., & Yuan, K.-H. (2017). New ways to evaluate goodness of fit: A note on using equivalence testing to assess structural equation models. *Structural Equation Modeling*, *24*(1), 148–153.
- Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713–732.
- Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York: Routledge.
- Muthén, B. (1993). Goodness of fit with categorical and other nonnormal variables. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 205–234). Newbury Park, CA: Sage.
- Muthén, L. K., & Muthén, B. O. (1998–2016). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
- Muthén, B. O., du Toit, S. H., Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Unpublished technical report.
- R Development Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78–90.
- Wicherters, J. M., & Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: An illustration using IQ test performance of minorities. *Educational Measurement Issues and Practice*, *29*(3), 39–47.
- 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 & Evaluation*, *12*(3), 1–26.
- Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, *30*(1), 165–200.
- Yuan, K.-H., & Bentler, P. M. (2004). On chi-square difference and Z tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological Measurement*, *64*, 737–757.

- Yuan, K.-H., & Chan, W. (2016). Measurement invariance via multigroup SEM: Issues and solutions with chi-square difference tests. *Psychological Methods, 21*(3), 405–426.
- Yuan, K.-H., Chan, W., Marcoulides, G. A., & Bentler, P. M. (2016). Assessing structural equation models by equivalence testing with adjusted fit indexes. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(3), 319–330.