Evaluating Equivalence Testing Methods for Measurement Invariance

Alyssa Counsell[1*], Robert A. Cribbie[2], & David B. Flora

York University, Toronto, ON

[1] ORCiD = 0000-0001-9449-6630; Twitter = @AlyssaCounsell

[2] ORCiD =0000-0002-9247-497X; Twitter = @rcribbie

This work was partially supported by the Social Sciences and Humanities Research Council (SSHRC) under Grant 435-2016-1057.

* Correspondence should be addressed to Alyssa Counsell who has moved to the Department of Psychology at Ryerson University, Toronto, ON, Canada, M5B 2K3 (e-mail: a.counsell@ryerson.ca)

Note: This paper is a previous version of the manuscript with the same name published in *Multivariate Behavioral Research*.

Abstract

Measurement Invariance (MI) is often concluded from a nonsignificant chi-square difference test. Researchers have also proposed using change in goodness of fit indices (ΔGOFs) instead. Both of these commonly used methods for testing MI have important limitations. To combat these issues, Yuan and Chan (2016) proposed using an equivalence test (EQ) to replace the chi-square difference test commonly used to test MI. Due to their concerns with the EQ's power, Yuan and Chan also created an adjusted version (EQ-A), but provide little evaluation of either procedure. The current study evaluated the Type I error and power of both the EQ and EQ-A, and compared their performance to that of the traditional chi-square difference test and ΔGOFs. The EQ for nested model comparisons was the only procedure that always maintained empirical error rates below the nominal alpha level. Results also highlight that the EQ requires larger sample sizes than traditional difference-based approaches or using equivalence bounds based on larger than conventional RMSEA values (e.g., > .05) to ensure adequate power rates. We do not recommend Yuan and Chan's proposed adjustment (EQ-A) over the EQ.

Evaluating Equivalence Testing Methods for Measurement Invariance

Psychological constructs like self-esteem and depression are not directly observable; instead, their measurement must be approximated by imperfect scales. Measuring constructs in this way has implications for assessing differences on the construct between populations. Specifically, researchers want to ensure that any observed group differences are a function of only true construct differences and not confounded with differences in how the construct is measured across groups. One way to determine whether an instrument measures the construct equivalently across multiple groups is by testing measurement invariance (MI) in the context of a multiple-group confirmatory factor analysis (CFA) for the individual items within a scale. Tools which allow researchers to test their instruments, scales, and composite measures for MI facilitate more accurate portrayals of the differences in psychological constructs across different groups. In fact, common tests of mean differences (e.g., *t*-test, analysis of variance) are biased if MI cannot be established for the scale of interest.

Measurement invariance is commonly tested using chi-square difference tests for comparing nested models with increasingly stricter across-group equality constraints on model parameters. Although MI testing involves model comparisons to examine the plausibility of invariance constraints, it is first necessary to establish adequate overall fit for the unconstrained multigroup CFA model. This first step of establishing adequate overall model fit is typically referred to as configural invariance (Horn & McArdle, 1992), in which each group has the same pattern of fixed and free parameters, but no parameters are constrained to be equal across groups (except for those needed to establish model identification). Once configural invariance has been established, there are further levels of MI that one could examine. These levels are typically tested in the sequence described below and if a preceding level is not satisfied, it is not appropriate to test further levels (Byrne, Shavelson, & Muthén, 1989; Millsap, 2011; Reise, Widaman, & Pugh, 1993). The second MI step, called metric invariance (Horn & McArdle, 1992) or weak invariance (Widaman & Reise, 1997), is met when factor loadings are equal across groups. If this condition is satisfied, then scalar invariance (Steenkamp & Baumgartner, 1998) or strong invariance (Meredith, 1993) can be tested, in which the intercepts in the regression of the observed indicators on the common factors are equal across groups. The last level of MI is strict invariance (Meredith, 1993), which states that the indicators' unique or error variances are equal across the groups. Either strict or scalar invariance is considered necessary for valid observed mean comparisons across different populations.

Testing these latter three levels of invariance involves placing across-group equality constraints on the parameters of interest (i.e., factor loadings, intercepts, or unique variances) such that the models corresponding to different levels of MI are nested models. With nested models, chi-square difference tests are typically used to determine whether these constraints produce a statistical worsening of model fit, that is, whether the equality constraints are not plausible, given the data. With the goal of finding support for MI, a researcher wants to obtain a nonsignificant chi-square difference to conclude that parameters do not differ across groups and thus the equality constraints cannot be rejected. However, using chi-square difference tests in this way constitutes a logical mistake for MI testing because a researcher's goal is to find support for the null hypothesis of equality across group parameters, not to find differences. In general, failing to reject a null hypothesis is not a statistically valid approach to determine similarity or equivalence. In fact, a number of researchers have argued that using chi-square difference tests for MI is problematic (e.g., Bentler, 1990; Browne & Cudeck, 1992; Cudeck & Henly, 1991; Kang, McNeish, & Hancock, 2016; MacCallum, Browne, & Sugawara, 1996). Recently, Yuan, Chan, Marcoulides, and Bentler (2016) and Yuan and Chan (2016) argued for using equivalence testing counterparts for the chi-square (for testing model fit/configural invariance) and chi-square difference statistics to avoid this logical mistake.

Equivalence Testing

Equivalence testing represents a category of logical and valid statistical procedures for researchers seeking statistical support for hypotheses of equivalence or similarity. Many univariate equivalence tests have been developed for evaluating means (e.g., Anderson & Hauck, 1983; Koh & Cribbie, 2013; Mara & Cribbie, 2012; Schuirmann, 1987; Wellek, 2010; Westlake, 1972) and correlations (e.g., Counsell & Cribbie, 2015; Goertzen & Cribbie, 2010). Less work has applied equivalence testing to multivariate procedures. One such reason may be the difficulty of establishing an appropriate equivalence interval, which refers to bounds such that effects contained therein are considered practically negligible or inconsequential (Rogers, Howard, & Vessey, 1993). Despite challenges associated with choosing an equivalence interval, equivalence testing represents a conceptually appropriate method for testing MI.

Equivalence Testing Methods for Measurement Invariance

Yuan and Chan (2016) outline two equivalence testing approaches which allow for a small amount of model misspecification (reflecting an equivalence interval) by incorporating a noncentrality parameter into the null hypothesis. The first test is used in lieu of the traditional chi-square statistic to assess the discrepancy between a sample covariance matrix and a model-implied covariance matrix and thus may be applied to testing configural invariance. The second test is the equivalence-based version of the chi-square difference test for nested model comparison and may be used for subsequent stages of invariance testing. Recall from above that it is assumed that the researcher's goal is to establish measurement invariance (i.e., similar measurement parameters across groups), and hence equivalence testing methods are appropriate.

Chi-square equivalence test

The first equivalence test proposed by Yuan et al. (2016) and Yuan and Chan (2016) evaluates model fit by allowing a small degree of model misspecification. While this test can be used in single-group CFA or structural equation modelling (SEM) generally, it is relevant to MI because it is used at the configural stage to test that the same factor structure fits well in each group. The traditional chi square statistic in SEM is calculated as:

$$T_{ML} = (N-1)F_{ML} \tag{1}$$

where N is the sample size and F_{ML} is the maximum-likelihood statistic giving the discrepancy between the sample covariance matrix (and mean vector) and model-implied mean and covariance structure. For a properly specified model, T_{ML} is distributed as χ^2 with degrees of freedom (*df*) equal to the total number of non-redundant elements in the sample covariance matrices and mean vectors minus the number of free parameters in the multiple-group model. Because T_{ML} is distributed as χ^2, it is often simply called the chi-square statistic. To create an equivalence test version of this statistic, Yuan and Chan (2016) modified the null hypothesis to H_0: $F_{ML0} > \varepsilon_0$ instead of the traditional $F_{ML0} = 0$, where F_{ML0} is the population counterpart of F_{ML} and ε_0 is a positive number that the researcher can tolerate for the size of misspecification, that is, the value for the equivalence bound. But because the null hypothesis is one-sided, the value for ε_0 is a single number rather than an interval. Under this slight misspecification, T_{ML} is distributed as a noncentral χ^2 statistic, with ε_0 used to calculate the corresponding noncentrality parameter, $\delta_0 = (N-1)\varepsilon_0$. With $c_\alpha(\varepsilon_0)$ as the left-tail critical value of the noncentral $\chi^2(\delta_0)$ at cumulative probability α, one rejects $H_0$when $T_{ML} \leq c_\alpha(\varepsilon_0)$. Rejection of the null hypothesis implies that the model misspecification is smaller than the pre-specified equivalence bound, ε_0. In this case, a researcher can conclude that any differences between the sample means and covariances and model-implied mean and covariance structure is trivial. In context of MI, a researcher would seek to reject H$_0$: $F_{ML0} > \varepsilon_0$ in each group to conclude that configural invariance is satisfied. Details on calculating ε_0 are provided below after describing the equivalence testing method for the chi-square difference test. The approach to test a single group model's model fit separately to conclude configural invariance is the approach recommended by Yuan and Chan (2016). However, a researcher could also examine a multigroup chi-square statistic calculated as:

$$T_{ML} = (N - K)F_{ML} = (N_1 - 1)F_{ML}^{(1)} + (N_2 - 1)F_{ML}^{(2)} + \cdots + \ (N_K - 1)F_{ML}^{(K)} \qquad (2)$$

for *K* groups, where *N*$_1$, *N*$_2$,…, *N*$_K$ are the group sample sizes and $F_{ML}^{(1)}$, $F_{ML}^{(2)}$, …, $F_{ML}^{(K)}$are the maximum-likelihood statistics giving the discrepancy between the sample covariance matrix (and mean vector) and model-implied mean and covariance structure within each group.
Note that T_{ML} retains all of the same properties specified above in the single-group case. One key distinction, however, is that for the purpose of nested model comparisons, one necessarily must estimate this multigroup model even if model fit is assessed separately in each group.

Chi-square difference equivalence test

The chi-square difference test calculates the difference between the T_{ML} statistics of two nested models, denoted $T_{bc} - \ T_b$, where the *b* subscripts refer to the baseline model and *bc* refers to baseline model with additional across-group equality constraints:

$$T_{bc} - \ T_b = (N - K)(F_{bc} - \ F_b)$$

$$= (N_1 - 1)(F_{bc}^{(1)} - F_b^{(1)}) + (N_2 - 1)(F_{bc}^{(2)} - F_b^{(2)}) + \cdots + \ (N_K - 1)(F_{bc}^{(K)} - F_b^{(K)}) \quad (3)$$

Like the T_{ML} statistic, $T_{bc} - T_b$ is distributed as χ^2, where the *df* for the chi-square difference test equals the difference between the *df* of the baseline model and the *df* of the constrained model. Once again, Yuan and Chan (2016) modified the null hypothesis from $H_0 \colon F_{bc0} - F_{b0} = 0$, where *Fbc0* and *Fb0* correspond to the population counterparts of F_{bc} and F_b, to $H_0 \colon F_{bc0} - F_{b0} > \varepsilon_0$ to allow a small degree of model misspecification. Here, $T_{bc} - T_b$ is distributed as a noncentral χ^2 statistic with the corresponding noncentrality parameter for multigroup models:

$$\delta_0 = (N - K)\varepsilon_0 \tag{4}$$

One rejects the null hypothesis when $T_{bc} - T_b \leq c_\alpha(\varepsilon_0)$, where $c_\alpha(\varepsilon_0)$ is defined above. A statistically significant result provides evidence that the equality constraints are plausible, given the data. Consequently, the researcher would conclude that metric, scalar, or strict invariance has been met depending on which MI stage is currently being tested.

What is an appropriate equivalence interval?

Because the value of ε_0 is crucial to Yuan and Chan's (2016) equivalence tests for MI, it is important to discuss how one chooses its value. In their paper, ε_0 represents the largest amount of model misspecification a researcher is willing to accept. The choice of a reasonable value of ε_0, however, is not immediately evident. Thus, Yuan and Chan (2016) and Yuan et al. (2016) relate ε_0 to the root mean square error of approximation (RMSEA) based on the work of Steiger (1988, 1998):

$$\varepsilon_0 = df(RMSEA_0)^2/K \tag{5}$$

where $RMSEA_0$ is the a priori value of RMSEA that a researcher is willing to accept as a reasonable amount of misspecification (e.g., .05, .08) and *df* is the model degrees of freedom in a single-group case (i.e., *K* = 1) or difference in *df* when comparing nested models. An important consideration is that for a given value of $RMSEA_0$, ε_0 changes depending on the model *df*, implying that a researcher cannot effectively choose a single value for ε_0 to be applied to multiple models like is typically done with a simple adoption of Hu and Bentler's (1999) fit index recommendations. As an aside, common guidelines for descriptive goodness of fit (GOF) (e.g., Hu & Bentler, 1999) also were not meant for use as strict cut-off values, but applied researchers typically use them as such.

Yuan and colleagues (2016) argued for a standard for choosing the value of ε_0 and proposed using MacCallum et al.'s (1996) RMSEA guidelines of .01, .05, .08, and .10 for excellent, close, fair, mediocre, and poor fitting models. They later noted, however, that using these guidelines to calculate values of ε_0 in the equivalence test results in too stringent an amount of model misspecification compared to using them with the traditional point-estimate null hypothesis. To address this issue, they created adjusted RMSEA values for use in calculating ε_0 and propose that these should be the new norm for cut-offs in the equivalence test (Yuan et al., 2016; Yuan & Chan, 2016). The exact details of this adjustment can be seen in Table 12 in Yuan and Chan (2016). They also provide a function to calculate the adjusted RMSEA from conventional values of .01, .05, .08, and .10 and have recently created the R package `equaltestMI` (Jiang, Mai, & Yuan, 2017) to facilitate using these tests.

Power and Type I Errors

An important consideration is that the language around a test statistic's performance changes when comparing equivalence tests to tests assessing parameter differences. Table 1 demonstrates how the concepts of Type I error and power differ across the traditional chi-square (hereby denoted TCS) and its equivalence testing (EQ) counterpart. When comparing substantive results using these different approaches, it can be confusing to use terms such as *Type I error* or *power* because the TCS and its EQ counterpart test different null hypotheses. Instead, we use language that reflects arriving at the same conclusion. Specifically, we refer to concluding invariance when population parameters are truly invariant as *correctly concluding invariance*, whereas we use the phrase *falsely concluding invariance* for concluding invariance when the population parameters are not invariant.

Change in Goodness of Fit Cut-Offs for MI

Due to the limitations of the TCS, some researchers have advocated instead using change in descriptive GOF indices between nested models to test MI (e.g., Chen, 2007; Cheung & Rensvold, 2002). Although there are many alternative fit indices, we focus on three: the RMSEA, the comparative fit index (CFI; Bentler, 1990), and McDonald's noncentrality index (MNCI; McDonald, 1989). The CFI and RMSEA were chosen because they are the most commonly reported fit indices in applied research (Jackson, Gillaspy, & Purc-Stephenson, 2009), due in part to their good statistical properties. The MNCI was included because research suggests that it performs well for comparing models in the context of MI testing (Cheung & Rensvold, 2002; Kang et al., 2016; Meade, Johnson, & Braddy, 2008).

Despite the large number of fit indices developed to reconcile problems with using the TCS alone, they have their own limitations. Yuan (2005) notes that many fit indices do not have known sampling distributions and do not have specific null hypothesis associated with them. Therefore, cut-offs for these GOF indices cannot be used like a traditional critical value in hypothesis testing. This criticism does not apply to all fit indices, though, as one can specify a null hypothesis for RMSEA (MacCallum et al., 1996) and software commonly reports a test of a null hypothesis that RMSEA < .05. Many researchers recommend avoiding strict adherence to popular GOF cut-offs (e.g., Marsh, Hau, & Wen, 2004; Steiger, 2007), but in the context of MI testing, a number of papers provide recommendations for GOF cut-offs based on simulation work. For example, Cheung and Rensvold (2002) recommended |ΔMNCI| ≤ .02 and |ΔCFI|≤.01 as indicative of equivalent fit across nested models whereas Chen (2007) advocated using cut-offs of |ΔCFI|≤ .005 and |ΔRMSEA|≤ .01.

Rationale for the Current Study

One limitation of Yuan and Chan's (2016) paper is a lack of empirical evaluation into why a researcher should use their adjusted RMSEA$_0$ value to calculate the equivalence bound, ε_0. Instead, they provide a theoretical justification stating that the probability of rejecting the null hypothesis of non-equivalence decreases as one proceeds to testing later stages of MI due to the sequential nature of the nested model comparisons. For example, with just two groups, the probability is raised to the fifth power to reach the end of the MI sequence: *p*(group 1 configural model fit)**p*(group 2 configural model fit)**p*(metric)**p*(scalar)**p*(strict invariance). While this sequence certainly results in decreased power and evidence that traditional RMSEA values may be too stringent, the extent to which Yuan and Chan's proposed adjusted RMSEA remedies this problem is unknown. Their equivalence tests were also not compared to recommended GOF cut-offs. Therefore, the current study evaluates the performance of Yuan and Chan's (2016) proposed equivalence tests for MI, including their adjusted RMSEA approach, and compares the results to those obtained using commonly used GOF cut-offs.

Method

A simulation study was used to collect probability rates for concluding invariance for Yuan and Chan's (2016) equivalence test (EQ) and the EQ using adjusted values of ε_0 (EQ-A) under conditions where population parameters are either invariant across groups or not (i.e., power or Type I error conditions, respectively, for the equivalence tests). The results from the EQ and EQ-A are also compared to results from a traditional chi-square (TCS) test with a statistically nonsignificant result (i.e., $p > \alpha$), as well as results from previously recommended ΔGOF cut-offs for CFI, RMSEA, and MNCI. At the configural stage, we used GOF cut-offs of .95 for CFI, .95 for MNCI, and values of the RMSEA that corresponded to what was used to calculate ε_0 (i.e., .05, .08, and .10). Cut-offs used for the ΔGOFs at the metric, scalar, and strict invariance stages were taken from Chen (2007). Consequently, invariance was concluded if the CFI of the more constrained model did not decrease by more than .005, the RMSEA did not increase by more than .01, and the MNCI did not decrease by more than .01. The simulation was run in R (R Core Team, 2016), and data generation and analysis were carried out using the *lavaan* package (Rosseel, 2012). Data were generated by specifying a population CFA model separately for each of two groups and then randomly sampling data that matched each group's model-implied covariance matrix and mean structure. Each study design cell included 5000 replications.

In all conditions, the population model included two latent variables that were correlated at .5, with no unique covariances among the indicators (except under specific conditions at the configural stage described below). A number of conditions were manipulated including: number of indicators (one condition with four indicators per factor and one with eight indicators per factor), size of standardized factor loadings (.5, .7, or .9 with standardized error variances corresponding to .75, .51, and .19, respectively), amount of population model misspecification, $F_{bc} - F_b$, sample size per group (100, 250, 500, 1000, or 2000), and whether the population models are characterized by MI (which is a power condition in the equivalence testing perspective) or not (Type I error condition for equivalence testing). Details of the amount of model misspecification in the Type I error conditions are given below.

The configural invariance (unconstrained multigroup CFA) model was estimated such that the latent variables in each group were standardized (for identification) and all other parameters were freely estimated. The metric invariance model allowed the variance of the latent variables to be freely estimated and the model was identified by fixing the first indicator's loading on each latent variable equal to one with across-group equality constraints on the rest of the indicators' loadings. The scalar invariance model freed the mean and variance of the first group's latent variables and included across-group equality constraints on all of the indicators' intercepts. Finally, the strict invariance model added group equality constraints on the error variances.

Probability of Correctly Concluding Invariance Conditions

To investigate the probability of correctly concluding invariance in Yuan and Chan's (2016) proposed equivalence tests, we examined two different conditions. In the first condition, all population model parameters were exactly equal in each of the two groups. In the second condition, the population models were not identical, but the differences between them were minute, such that any parameter differences were small enough to be considered trivial. Specifically, population model misspecification was set at 10% of the equivalence bound, ε_0 based on RMSEA$_0$ = .10, which was always smaller than the amount of tolerable misspecification of ε_0 calculated from RMSEA$_0$ of .05, .08, and .10. We focus on ε_0 in RMSEA$_0$ units instead of raw F_{ML} units for two reasons: We believe that RMSEA units are more intuitive to most readers than F_{ML} units, and the value of ε_0 changes based on the model *df*, so discussing ε_0 calculated from common RMSEA$_0$ units facilitates comparing probabilities of concluding invariance across the four- and eight-indicator models.

Because rates of correctly concluding invariance for a later stage of MI are calculated only when the previous stages have been satisfied, the discrepancy between population models as one moves through the stages is not additive for the conditions with small group differences in parameters (which is also true for population noninvariance conditions described below). Using scalar invariance as an example, rates of correctly concluding invariance at this stage are calculated with a small degree of model misspecification in an intercept, but the small degree of misspecification in loadings previously in the metric invariance stage would no longer be present in our population models. Regardless of whether the groups' population parameters were identical or had small differences, rates of correctly concluding invariance for the metric, scalar, and strict invariance stages all take into consideration whether the previous stages have been satisfied (e.g., rates at the metric stage would be calculated as *p*(configural group 1)**p*(configural group 2)**p*(metric) because metric invariance is assessed only after acceptable model fit in each group has been demonstrated). The implication of calculating probabilities this way is that the rates at a given stage of MI are not true representations of a statistical test's rates of correctly concluding invariance in and of itself, but of a test's rates within the MI sequence.

Probability of Incorrectly Concluding Invariance Conditions

We also sought to investigate Type I error control in Yuan and Chan's (2016) proposed equivalence test. Once again, to avoid language confusion surrounding Type I error across the different methods, we refer to probabilities under these conditions as rates of incorrectly concluding invariance. This condition was created such that the amount of model misspecification in the population model was $F_{ML0} = \varepsilon_0$ for each group separately at the configural stage, and $F_{bc0} - F_{b0} = \varepsilon_0$ in the metric, scalar, and strict invariance stages. This manipulation is consistent with previous equivalence testing literature evaluating an equivalence test's Type I error rates (e.g., Cribbie et al., 2004; Rogers et al., 1993; Schuirman, 1987), because one would expect the largest number of incorrect rejection rates when the population effect under study is exactly at the equivalence bound. Measuring rates of incorrectly concluding invariance in this way means using a slightly different null hypothesis than that of Yuan and Chan (2016). Their tests' null hypotheses were $F_{ML0} > \varepsilon_0$ and $F_{bc0} - F_{b0} > \varepsilon_0$, whereas the null hypotheses we use (which are more consistent with the equivalence testing literature) are $F_{ML0} \geq \varepsilon_0$ and $F_{bc0} - F_{b0} \geq \varepsilon_0$. Although the distinction is subtle, the implication is that Yuan and Chan consider ε_0 to be the largest amount of model misspecification one is willing to tolerate for concluding invariance, whereas we consider ε_0 to be the smallest amount of model misspecification that a researcher would deem a meaningful difference and would, therefore, indicate noninvariance. Note that in our results section, we refer to the amount of population model misspecification in RMSEA$_0$ units equal to those in calculating ε_0 instead of $F_{bc0} - F_{b0}$ units for the same reasons noted earlier.

Similar to the condition of correctly concluding invariance, the amount of model misspecification when testing rates of incorrectly concluding invariance at a later stage included only the misspecification at that level, such that invariance was met in the previous stages. For example, when assessing the rates of incorrectly concluding scalar invariance, the amount of model misspecification $F_{bc0} - F_{b0}$ affected only the intercepts and not any of the model parameters implicated in configural or metric invariance.

To violate configural invariance, unique or error covariances were added to different pairs of indicators in each group. In group one, a covariance was added between the uniquenesses of the first indicators on each factor and in group two, a unique covariance was added between the second indicators on each factor. The magnitude of the covariances varied by population measurement model, factor loading magnitude, and amount of model misspecification. To violate metric, scalar, or strict invariance, we tested two different conditions. In the first, a single parameter (loading, intercept, or error variance) was noninvariant across the two groups; in the second, 25% of the parameters were noninvariant across the two groups. In both cases (i.e., single noninvariant loading or 25% noninvariance loadings), however, the amount of model level misspecification ($F_{bc0} - F_{b0}$) was equivalent. The nominal Type I error rate (α) for equivalence tests (i.e., rate of falsely concluding invariance) was set to .05 for all investigated conditions and empirical rates were considered acceptable if they fell within Bradley's (1978) liberal bounds of α +/- .5α.

Results

Although we have results for all of the conditions outlined in the methods section above, we restrict the results presented here to stages of invariance that involve model comparison approaches. This decision was due to both space constraints and to allow for comparisons of methods with practical utility; few applied researchers rely on the TCS to assess model fit at the configural invariance stage, but many still use the TCS difference test at later invariance stages. Full tables of results are available on the authors' Open Science Framework (OSF) page at https://osf.io/49fkd/.Because the same statistical technique is used at the metric, scalar, and strict stage, rates of concluding invariance at the scalar and strict invariance levels are not needed to evaluate the statistical method itself, but instead reflect the performance of the method raised to a power within the MI sequence. Accordingly, we focus on the numeric results for metric invariance.

Nonconvergence

Nonconvergence was not an issue for any of the models tested. Specifically, the highest rate of nonconvergence was only 1.68% in the condition where population factor loadings were not invariant with the smallest sample size tested (i.e., 100 per group) and factor loadings of .5 in the four-indicator model. Rates of nonconvergence for sample sizes of 250 per group were negligible (e.g., between 0 and 0.2%). In replications where the model failed to converge, statistics were not recorded and rates of concluding invariance were calculated with the denominator adjusted by removing these problematic results (i.e., 5000 – number of nonconverged replications).

Rates of Incorrectly Concluding Invariance

Note that rejection rates under these conditions represent Type I error rates for the equivalence-based procedures and Type II error rates for TCS. Figure 1 illustrates the metric invariance performance of all tests in the four-indicator model, with average factor loading parameters (i.e., .7). The top row includes information when both the amount of population model misspecification (based on RMSEA_0) and ε_0 were set at .05, whereas the bottom row includes information when both the amount of population model misspecification (RMSEA_0) and ε_0 were set at .10. Figure 2 illustrates the performance of the various tests in the metric invariance condition with population model misspecification of $\text{RMSEA}_0 = .08$ in the eight-indicator model. The top row includes results from low factor loading conditions (.5) and the bottom row includes results from high factor loading conditions (.9). All figures include the rates of falsely concluding invariance at the metric stage, ignoring whether configural invariance was met. We also collected rates of falsely concluding invariance for each of the tests only for the replications where configural invariance was satisfied. These rates were based on fewer replications (i.e., *p*(correctly concluding configural invariance)*number of replications), but the results are almost identical to those discussed.

Equivalence Based Chi Square Difference Tests. The EQ and EQ-A adopt the same approach but use different equivalence bounds to calculate the noncentrality parameter. The EQ demonstrated accurate error rates across testing metric, scalar, or strict invariance, and these rates were not affected by sample size, measurement model, factor loading magnitude, single parameter loading noninvariance versus multiple factor loading noninvariance, or magnitude of model misspecification. The EQ-A had rates of falsely concluding invariance around .50 under all conditions except when the sample size was 2000 per group. With *n* = 2000 per group, the rates decreased as RMSEA_0 decreased. Under the conditions tested, the adjustment provided an overcorrection to combat reduced power (i.e., inflated Type I error rates).

Traditional Chi-Square Difference Test. The TCS difference test had inflated rates of falsely concluding invariance at all stages. At the metric invariance stage, the TCS had rates as high as .81 at smaller sample sizes based on population model misspecification at RMSEA_0 = .05. As expected, these rates decreased as sample size increased, eventually reaching zero, where the TCS difference test obtained high power to detect *non-invariance*. This pattern of results was not impacted by measurement model, factor loading magnitude, or whether the noninvariance was driven by a single factor loading or multiple factor loadings.

***Goodness of fit cut-offs*.** Rates of incorrectly concluding invariance using ΔGOF were also high across the different stages of invariance. Results using ΔRMSEA were similar to the TCS, but the rates of falsely concluding invariance did not decrease as rapidly with increasing sample size, particularly with model misspecification at RMSEA_0 = .05 or .08. The ΔCFI produced rates of falsely concluding invariance which depended on all of the investigated conditions except for whether the noninvariance was driven by a single noninvariant parameter or 25% noninvariant parameters. Rates of falsely concluding invariance generally increased as the magnitude of the factor loadings increased and were highest with the smallest amount of model misspecification (e.g., RMSEA_0 = .05). These conditions also interacted, such that rates of falsely concluding invariance were lower as the amount of model misspecification increased and typically close to zero with the RMSEA_0 = .08 and .10 for medium to large sample sizes. The same pattern of results occurred for comparing the two measurement models, whereby rates of falsely concluding invariance were higher in the eight-indicator model at RMSEA_0 = .05, but lower in the eight-indicator model with RMSEA_0 = .08 or .10. The ΔMNCI's rates of falsely concluding invariance were unaffected by factor loading magnitude or single versus multiple loading invariance. In conditions with RMSEA_0 = .05, the MNCI's rates of concluding invariance increased with sample size whereas they stayed constant around .5 with RMSEA_0 = .08, and decreased with sample size with RMSEA_0 = .10.

Probability of Correctly Concluding Invariance

Note that rates under these conditions represent power rates for the equivalence-based procedures. Before discussing probabilities of correctly concluding invariance, it is important to note that different approaches' rejection rates are not fairly comparable unless accurate rates of falsely concluding invariance are observed under similar conditions (i.e., same measurement model and factor loading magnitude).

Figures 3 demonstrates power rates for the EQ test using different equivalence intervals (i.e., based on RMSEA_0= .05, .08, and .10) as well as the rates of correctly concluding invariance using the TCS. The left column illustrates rates when group population models are identical (i.e.,, all parameters are equal) and the right column illustrates rates when group population models have slight differences in population loadings but are still within the equivalence bounds. The top row demonstrates the performance of the tests without the impact of MI sequence whereas the bottom row demonstrates results at the metric invariance stage (i.e., takes into consideration the configural invariance rates). The plots do not include the EQ-A or GOF indices due to ceiling effects, but full tables of results are included on the OSF site.

Equivalence-based chi square difference tests. With no model misspecification or small differences in parameters (i.e., that fall below the equivalence bound), the EQ and EQ-A both demonstrate the appropriate statistical relationships between sample size and power across all conditions; that is, rates of correctly concluding invariance increased as sample size increased. Rates for the EQ-A were significantly higher than the EQ, but this result occurred because the EQ-A falsely concluded invariance in approximately 50% of replications (with some variability by condition). Power rates for the EQ were low at small sample sizes (100 or 250 per group) using an equivalence interval based on RMSEA_0 = .05, but they increase rapidly as ε_0 or N increases. The pattern of results was the same across the conditions with identical population model parameters and those with small differences in parameters, but, as expected, power rates were higher when population parameters were identical.

Traditional chi-square difference test. With no population model misspecification, rates for concluding invariance using the TCS do not change as a function of sample size. This result is expected because with no model misspecification, one would expect the rates to stay constant at 1-α (where α is the nominal Type I error rate for the TCS), i.e., .95. These rates are shown in the upper left graph in Figure 3. Rates are not .95 in the bottom left graph because the rates are impacted by position within the MI sequence (i.e., *p*(configural group 1)**p*(configural group 2)**p*(metric)), which should equal approximately .86. With inconsequential levels of population model misspecification, rates of correctly concluding invariance using the TCS decrease as sample size increases. In other words, the TCS demonstrated an inappropriate relationship between rates of correctly concluding invariance and sample size.

Goodness of fit cut-offs. Ceiling effects (i.e., rates of concluding invariance very close to or at 1.00) were observed for the rates of falsely concluding invariance for the ΔGOF indices in most of the tested conditions. This result is unsurprising, given that the tests concluded invariance under similar conditions when the population models were misspecified. When the groups' population models were identical, the rates of concluding invariance using a ΔRMSEA or ΔCFI were generally higher in the eight-indicator model compared to the four-indicator model, whereas they were generally higher in the four-indicator model using ΔMNCI. Rates of concluding invariance increased as sample size increased, but even at the smallest sample size (100 per group), rates were close to .6 or .7. Again, the ΔCFI interacted with factor loading reliability such that it tended to conclude invariance more often with higher reliabilities.

In conclusions, simulation results suggest that the EQ is the only test that demonstrates valid statistical properties across all of the conditions tested.

Applied Example

To demonstrate the equivalence testing methods proposed by Yuan and Chan (2016) in comparison with other popular methods (i.e., TCS and ΔGOF indices), we obtained data from a large scale personality testing website (http://personality-testing.info/), specifically, responses to the Generic Conspiracist Beliefs Scale (GCB; Brotherton, French, & Pickering, 2013). This scale includes 15 five-point Likert-type items measuring the degree to which individuals endorse five types of conspiracies, each represented as a latent variable in a five-factor model. Each conspiracy factor has three items that load solely on its factor (i.e., there are no cross loadings).

The conspiracy factors include *government malfeasance*, *extra-terrestrial cover-up*, *malevolent global conspiracies*, *personal well-being*, and *information control*. For the current example, we test MI on the scale using sex (male vs. female) as the grouping variable to determine whether the scale functions equivalently across sex. The dataset included 2359 participants, 1222 men and 1137 women. There were no missing data on any of the items on the GCB within either group. Data and R code to reproduce this analysis are made available at https://osf.io/49fkd/.

For each of the MI steps, we compared multiple-group CFA models estimated using maximum likelihood (ML) with the *lavaan* package (Rosseel, 2012) in R (R Core Team, 2016). Given that the item response variables are ordinal with five categories, Rhemtulla, Brosseau-Liard, and Savalei (2012) suggest that the variables may be treated as continuous rather than categorical, thus justifying fitting the CFA models to observed means and covariances using ML. While other estimation techniques (e.g., based on polychoric correlations) may be more appropriate, the equivalence testing methods have not been fully developed for categorical variables and the purpose of this section is to demonstrate how substantive conclusions may differ according to choice of MI testing procedure.

Statistical results for various MI stages are presented in Table 2 for each of the TCS, EQ, EQ-A, and three ΔGOF indices, namely CFI, RMSEA, and MNCI.

Results Using the Equivalence Testing Methods for MI

The equivalence testing R functions provided in Yuan and Chan (2016) use information easily obtained from any SEM software. However, the functions do not provide *p* values for the equivalence tests, and so we modified them to provide *p* values and allow users to choose the value of RMSEA_0 to calculate a single equivalence bound; our modified function is available on the OSF site with the rest of the applied example materials. One could also use the `equaltestMI` package (Jiang et al., 2017) to run through the full MI sequence employing equivalence tests across a range of RMSEA$_0$ values. Although the functions in the `equaltestMI` package include multiple RMSEA cut-offs for calculating ε_0, we believe that it is more appropriate for researchers to choose an appropriate level of model misspecification a priori. For this example, we use RMSEA$_0$ = .08 to calculate ε_0 and its corresponding noncentrality parameter δ_0 for each model comparison. Note that the EQ and EQ-A represent the same statistical test, but use different values for ε_0 and δ_0 because RMSEA$_0$ is adjusted for EQ-A (to values between .086 and .09, depending on the model).

Yuan and Chan (2016) recommend demonstrating adequate model fit separately for each group to establish configural invariance (rather than testing a single multiple-group model without equality constraints), so Table 2 includes separate EQ-based model fit statistics for men and women at the configural stage (see Yuan et al., 2016 for using the EQ to test model fit). Further stages of invariance testing involve using the equivalence test version of the chi-square difference test. Because the EQ and EQ-A included statistically significant results for all of the invariance stages except for strict invariance, results suggest that the five-factor GCB meets scalar invariance across sex.

Results Using the Traditional Chi Square Test

If a researcher relied on the TCS alone, configural invariance for the five-factor model would not be concluded because statistically significant χ^2 fit statistics were obtained for both women and men (see Table 2)**.** Without re-specifying the five factor model described above, it would be inappropriate to further test the invariance of loadings, intercepts, or unique variances. In practice, however, it is rare for a researcher to rely exclusively on the TCS to assess overall model fit. Because the CFI and RMSEA indicated adequate model fit within each group, we examined further levels of invariance using the TCS difference testing approach.[1] Table 2 indicates that a nonsignificant chi-square difference statistic was obtained for comparing the configural invariance model with the metric invariance model, suggesting that group equality constraints on factor loadings do not significantly worsen model fit. However, placing further group equality constraints on intercept parameters produces a significant chi-square difference statistic, thus rejecting scalar invariance. Therefore, using the TCS approach, one would conclude that the GCB satisfies metric but not scalar invariance.

Results Using the ΔGOF Indices

Cut-offs for demonstrating adequate overall model fit were .95 for both CFI and MNCI and .08 for RMSEA. Cut-offs for substantial change in model fit were .005 for CFI and .01 for RMSEA and MNCI. At the configural stage, adequate model fit was satisfied according to CFI and RMSEA, but not MNCI. Thus, if basing model fit on MNCI only, configural invariance could not be concluded. However, for the sake of the demonstration, we will assume that researchers would use other information to support that the five-factor structure fits the model relatively well in each group and use the ΔMNCI to test other levels. Using ΔGOF, Table 2 shows that one would conclude that scalar invariance is satisfied according to ΔCFI, whereas strict invariance is satisfied using ΔRMSEA and ΔMNCI.

This demonstration illustrates that a researcher can arrive at different substantive conclusions about the degree of MI depending upon the method used, with the equivalence test approaches supporting scalar invariance by sex for the five-factor model of the GCB. Table 2 includes a shaded region to highlight the stage of invariance concluded with each method.

[1] In fact, it is common for researchers to establish configural invariance using descriptive GOF measures, but then proceed to assess metric, scalar, and strict invariance using TCS difference tests.

Discussion

The most popular approaches to testing measurement invariance, namely using nonsignificant chi-square difference tests or using ΔGOF cut-offs, each have important limitations. Because traditional chi-square tests use a null hypothesis of no difference, MI is aligned with the null hypothesis, and failing to reject the null hypothesis does not provide evidence that invariance has been satisfied. That is, failing to reject H_0 does not imply that the model parameters are equivalent across the groups, only that there is not enough information to reject the null hypothesis that model parameters are identical across groups. This situation is particularly problematic because the chi-square statistic often has high power to detect minor model misspecification with the larger sample sizes typically used in CFA or SEM (Byrne, 2008; MacCallum et al., 1996; Saris, Satorra, & van der Veld, 2009). One criticism of using ΔGOFs like ΔCFI or ΔRMSEA is the lack of consistent cut-offs values from simulation research. This limitation is unsurprising, however, because most GOFs are primarily descriptive with no known sampling distributions. Instead of relying on these procedures for testing MI, one could use a statistical tool which allows a small degree of model misspecification and allows rejection of a null hypothesis to provide evidence of invariance; Yuan and Chan's (2016) equivalence tests for MI do exactly that.

Performance of the Equivalence Tests for MI

In the majority of conditions tested in our simulation study, the EQ demonstrated accurate empirical Type I error rates and reasonable power. We argue that the EQ test in and of itself does not have problems with power, but due to the sequential testing of the MI stages, power to establish model equivalence becomes lower as one progresses through the sequence. Yuan and Chan (2016) and Yuan et al. (2016) discussed this issue, but without any empirical evaluation of the test's power. They provide functions that make an automatic adjustment of conventional RMSEA values of .01, .05, .08, and .10 and use these new adjusted values to calculate the population equivalence bound, ε_0. The current simulation demonstrated that this adjustment substantially impacts the rates of falsely concluding invariance compared to a population model generated under a typical user-specified RMSEA$_0$ (i.e., without the adjustment). In fact, rates of falsely concluding invariance in the simulation were sometimes ten times higher using the EQ-A instead of the regular EQ test. It is important to note that technically the EQ-A tests different a null hypothesis than the EQ. The statistical method used for both the EQ and EQ-A is the same, but the RMSEA$_0$ used to calculate ε_0 in the EQ-A is different from that of the EQ. The implication then is that the simulation results for the EQ-A are not true Type I error rates. The issue, however, is that it is unclear to an applied researcher conducting an equivalence test exactly how much the EQ-A is adjusting RMSEA$_0$ (and by extension, ε_0). For example, a researcher may feel comfortable using ε_0 calculated from RMSEA$_0$ = .08, but the adjusted RMSEA$_0$ might be closer to .10. The point of the simulation is to show that the EQ-A makes a large adjustment without much theoretical justification. If power is a concern as Yuan and Chan (2016) discuss, a researcher should choose a larger value RMSEA$_0$ rather than relying on a function to inflate it mechanistically.

Using Change in Goodness of Fit Indices

Because using a change in fit indices has been recommended as a remedy for the limitations of the traditional chi-square difference test, we included common recommendations for cut-offs based on Chen (2007). Before discussing their performance in the simulation, we note that comparing their rates of falsely concluding invariance to those of the EQ is somewhat unfair because we used a single cut-off instead of a value that may more accurately correspond to the amount of model misspecification. Yet, relying on a single cut-off is typically how applied researchers use a ΔGOF. Further, the primary goal of our study was to assess the performance of the EQ rather than test the sensitivity of GOF cut-offs. Given this purpose, it would have been excessive to include multiple cut-off levels for each of the three fit indices used in the current simulation study. Nonetheless, our results still contribute to the literature on the use of different cut-off points. For example, some researchers have recommended using ΔCFI (e.g., Chen, 2007; Cheung & Rensvold, 2002), while more recent research advocates against it because ΔCFI was found to fluctuate with factor reliability (Kang et al., 2016). Our simulation study supports Kang et al. because with high factor loadings, we found inflated rates of incorrectly concluding invariance, holding all else constant.

To conclude, it is difficult to directly compare the ΔCFI or ΔMNCI methods to the EQ method because the recommended cut-offs did not correspond well to any of the values of ε_0 tested in our simulation. In theory, one could change the cut-off to accommodate more or less model misspecification like is done with an equivalence interval, but it may be more difficult for researchers to determine a meaningful difference. Choosing a population value of the RMSEA to calculate ε_0, on the other hand, may be more intuitive for applied researchers because there is a more established literature about a range of recommended RMSEA values (e.g., MacCallum et al.'s (1996) "excellent" to "poor" fit).

Choosing an Equivalence Interval

Another important topic with equivalence testing is choosing an appropriate value for an equivalence interval or bounds. As equivalence testing was developed in the pharmaceutical field, there are standardized methods for determining whether two drugs have equivalent effects. These standards work well because different drugs can be compared using the same criteria (e.g., peak plasma level). In psychology, however, equivalence testing could be used for a variety of effect estimates beyond simple mean comparisons. Even for mean comparisons alone, it is likely not appropriate to use the same interval across different psychological scales. While a common criticism is that setting the interval introduces subjectivity because a researcher may choose any value that she likes, this criticism is not valid if the researcher's choice is theoretically justified. Rogers and colleagues (1993) noted that "as with any statistical analysis, equivalency procedures must involve thoughtful planning by the investigator" (p. 564). As long as the researcher chooses an interval before collecting data, and the value is appropriate for the research problem being addressed, the researcher is in no way introducing bias. For this reason, we argue that researchers should choose an RMSEA value to calculate the equivalence bounds (ε_0) before looking at the results to avoid choosing the value that produces the desired result; study preregistration can be effective for this purpose.

Within the context of Yuan and Chan's (2016) EQ, choosing an equivalence interval is less burdensome than for other types of equivalence tests because the value of RMSEA is the only real choice a researcher must make. In fact, the functions provided by Yuan and Chan do not even require the researcher to specify a level of the RMSEA and instead provide results at each of MacCallum et al.'s (1996) recommended cut-off points for levels of model fit. For use in equivalence testing, conventional RMSEA values such as .01 or .05 are generally too strict in practice, particularly with sample sizes that are considered small within the context of SEM. This issue was discussed by Yuan and Chan as well as Yuan et al. (2016), which led them to propose an adjusted RMSEA. As discussed above, we do not believe that the EQ-A is well justified (neither theoretically nor based on our simulation results). While we do not recommend using Yuan and Chan's EQ-A method, power for the EQ is an important consideration for applied researchers.

Limitations and Future Directions

As with any study, this work has some limitations. The measurement models chosen were meant to reflect models commonly encountered with psychological research, but any number of other models could be investigated; for instance, much larger models are estimated if a researcher uses scales with a large number of items. Furthermore, categorical item-level data typically requires a different estimation procedure for CFA models (e.g., one based on polychoric correlations) or those used in the item response theory (IRT) framework. The benefits of equivalence testing for MI in CFA models readily applies to differential item functioning or differential test functioning in IRT models, but different fit statistics are used, and therefore the equivalence interval would need to be reconsidered.

Readers should also be aware that the simulation included data that were all from a multivariate Gaussian distribution so that traditional maximum likelihood estimation was appropriate. In CFA or SEM models for variables with other distributions, many alternative estimators and adjustments for model fit statistics have been developed. Some examples include the Satorra-Bentler scaling correction (Satorra, 1992; Satorra & Bentler, 2001), weighted least squares methods (WLS; e.g., Browne, 1984; Muthén, 1984; Muthén & Satorra, 1995), and modified WLS methods (e.g., Muthén, du Toit, & Spisic, 1997; Yuan & Bentler, 1997; 1999). Incorporating equivalence testing into models using these methods would be beneficial, particularly as data in psychology are often not normally distributed and many CFA models in psychology are used within the context of scale validation (Jackson et al., 2009). In fact, our empirical example demonstrates the need for this extension. When extending Yuan and Chan's (2016) work to alternative estimators, calculation of ε_0 based on the same noncentral χ^2

distribution used for continuous data may not always be appropriate, so creating an appropriate equivalence interval for use with these estimation techniques may need additional work.

Another area for future research is to compare the results of the EQ to Bayesian methods for MI testing. Incorporating prior information into the model if previous invariance testing had occurred for the same groups on the same scale could also provide benefits for the EQ's power. A complication for comparing equivalence testing and Bayesian methods (which allow for finding evidence in favour of the traditional null hypothesis) is that information about where the effect falls relative to ε_0 is treated differently in the two methods. That is, while one could set a prior distribution around an effect of zero (on model or individual parameter differences), doing so will necessarily have an impact on the estimation of the effect. Equivalence tests, on the other hand, use only the data to estimate the effect. Creating a blended approach between the two based on incorporating an interval in the null hypothesis (e.g., Morey & Rouder, 2011) could take advantage of the best of both methods. Because Bayesian estimators do not rely on the same assumptions about distribution shape, incorporating them into the EQ also avoids relying on the noncentral χ^2 distribution.

Conclusion

Equivalence testing procedures are appropriate when a researcher seeks to demonstrate invariance. Based on the results previously described, we recommend that researchers adopt the original equivalence tests described in Yuan and Chan (2016) and Yuan et al. (2016) for MI testing. Their EQ versions of the chi-square and chi-square difference tests maintained accurate Type I error rates (i.e., rates of incorrectly concluding equivalence) and demonstrated statistically valid properties pertaining to power (i.e., rates of correctly concluding equivalence). We do not recommend using the EQ-A outlined in Yuan and Chan (2016) and Yuan et al. (2016)

because we believe that it is not theoretically justified and the degree to which it increases the researcher-specified equivalence interval may not be immediately evident to applied researchers. Based on simulation results, we also do not recommend using any of the commonly recommended cut-offs for ΔGOF indices in lieu of the EQ. While GOF indices sometimes perform better than the traditional chi-square test statistic, they do not demonstrate the valid statistical properties of the EQ.

References

Anderson, S., & Hauck, W. W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Statistics and Communications-Theory and Methods, 12,* 2663-2692. doi**:**10.1080/03610928308828634

Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological bulletin, 107,* 238.doi: 10.1037/0033-2909.107.2.238

Blanca, M. J., Arnau, J., Lopez-Montiel, D., Bono, R., & Bendayan, R. (2011) Skewness and kurtosis in real data samples. *Methodology, 92*, 78-84. Doi: 10.1027/1614- 2241/a000057

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31,* 144- 152.doi: 10.1111/j.2044-8317.1978.tb00581.x

Brotherton, R., French, C. C., & Pickering, A. D. (2013). Measuring belief in conspiracy theories: The Generic Conspiracist Beliefs Scale. *Frontiers in psychology, 4,* 1-15. http://dx.doi.org/10.3389/fpsyg.2013.00279

Browne, M. W. (1984). Asymptotically distribution free methods in the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37*, 127–141.doi: 10.1111/j.2044-8317.1984.tb00789.x

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

Bruder, M., Haffke, P., Neave, N., Nouripanah, N., & Imhoff, R. (2013). Measuring individual differences in generic beliefs in conspiracy theories across cultures: Conspiracy Mentality Questionnaire. *Frontiers in psychology, 4,* 225. https://doi.org/10.3389/fpsyg.2013.00225

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

Byrne, B. M. (1989). Multigroup comparisons and the assumption of equivalent construct validity across groups: Methodological and substantive issues. *Multivariate Behavioral Research, 24,* 503–523.doi: 10.1207/s15327906mbr2404_7

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

Chen, F. F. (2007). Sensitivity of goodness of fit indices to lack of measurement invariance. *Structural equation modeling, 14,* 464-504.doi: 10.1080/10705510701301834

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

Counsell, A., & Cribbie, R. A. (2015). Equivalence tests for comparing correlation and regression coefficients. *British Journal of Mathematical and Statistical Psychology*, *68*, 292-309. doi: 10.1111/bmsp.12045.

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

Darwin, H., Neave, N., and Holmes, J. (2011). Belief in conspiracy theories. The role of paranormal belief, paranoid ideation and schizotypy. *Pers. Individ. Dif.* 50, 1289– 1293. doi: 10.1016/j.paid.2011.02.027

Goertzen, J. R., & Cribbie, R. A. (2010). Detecting a lack of association: An equivalence testing approach. *British Journal of Mathematical and Statistical Psychology, 63*, 527-537. doi:10.1348/000711009X475853

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*, 117–144.doi: 10.1080/03610739208253916

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

Jackson, D. L., Gillaspy, J. A., Jr., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods, 14*, 6–23. doi: 10.1037/a0014694

Jiang, G., Mai, Y., & Yuan, K. H. (2017). equaltestMI: Examine measurement invariance via equivalence testing and projection method. R package version 0.1.0. https://CRAN.R-project.org/package=equaltestMI

Kang, Y., McNeish, D. M., & Hancock, G. R. (2016). The role of measurement quality on practical guidelines for assessing measurement and structural invariance. *Educational and Psychological Measurement*, *76*, 533-561. doi:10.1177/0013164415603764

Koh, A., & Cribbie, R. (2013). Robust tests of equivalence for k independent groups. *British Journal of Mathematical and Statistical Psychology, 66*, 426-434. doi: 10.1111/j.2044-8317.2012.02056.x

MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological methods*, *11*, 19-35.doi: 10.1037/1082-989X.11.1.19

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.

Mara, C. A., & Cribbie, R. A. (2012). Paired-samples tests of equivalence. *Communications in Statistics-Simulation and Computation*, *41*, 1928-1943. doi: 10.1080/03610918.2011.626545

Marsh, H.W., Hau, K.T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indices and dangers in overgeneralizing Hu and Bentler's findings. *Structural Equation Modeling, 11*, 320-41.doi: 10.1207/s15328007sem1103_2

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

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93,* 568 -592. doi: 10.1037/0021-9010.93.3.568.

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

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156-166. doi: 10.1037/0033-2909.105.1.156

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

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods, 16,* 406–419. doi:10.1037/a0024377

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

Muthén, B., du Toit, S. H.C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes (Unpublished manuscript).

Muthén, B. O.& Satorra, A. (1995). Technical aspects of Muthén's LISCOMP approach to estimation of latent variable relations with a comprehensive measurement model. P*sychometrika*, *60*, 489–503. doi:10.1007/BF02294325

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

RStudio Team (2016). RStudio: IntegratedDevelopment for R. RStudio, Inc., Boston, MA. URL http://www.rstudio.com/.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*, 552–566. doi: 10.1037/0033-2909.114.3.552

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

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113*, 553-565. doi: 10.1037/0033-2909.113.3.553

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48* (2), 1–36. Retrieved from http://www.jstatsoft.org/v48/i02

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. doi: 10.1080/10705510903203433

Satorra, A. (1992). Asymptotic robust inferences in the analysis of mean and covariance structures. *Sociological Methodology*, *22*, 249–278.

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

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*, 657-680. doi: 10.1007/BF01068419

Steiger, J. H. (1989). EzPATH: Causal modeling. Evanston, IL: SYSTAT.

Steiger, J. H. (1998). A note on multiple sample extensions of RMSEA fit index. *Structural Equation Modeling*, *5*, 411–419. doi: 10.1080/10705519809540115

Steiger, J.H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences, 42,* 893-98.doi: 10.1016/j.paid.2006.09.017

Wellek, S. (2010). Testing statistical hypotheses of equivalence and noninferiority (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Westlake, W. J. (1972). Use of confidence intervals in analysis of comparative bioavailability trials. *Journal of Pharmaceutical Sciences, 61* (8), 1340–1341. doi: 10.1002/jps.2600610845

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance abuse domain. In K. J. Bryant, *Alcohol and substance use research* (pp. 281-324). Washington, DC: APA.

Yuan, K. H. (2005). Fit indices versus test statistics. *Multivariate behavioral research*, *40*, 115-148. doi:10.1207/s15327906mbr4001_5

Yuan, K. H., & Bentler, P.M. (1997). Mean and covariance structure analysis: Theoretical and practical improvements. *Journal of the American Statistical Association, 92*, 767–774. doi: 10.1080/01621459.1997.10474029

Yuan, K. H., & Bentler, P. M. (1999). F tests for mean and covariance structure analysis. *Journal of Educational and Behavioral Statistics, 3*, 225–243.doi: 10.3102/10769986024003225

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. doi:10.1177/0013164404264853

Yuan, K. H., & Chan, W. (2016). Measurement invariance via multigroup SEM: Issues and solutions with chi-square-difference tests. *Psychological methods*, *21*, 405-426. doi: 10.1037/met0000080

Yuan, K. H., Chan, W., Marcoulides, G. A., & Bentler, P. M. (2016). Assessing structural equation models by equivalence testing with adjusted fit indices. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 319-330. doi: 10.1080/10705511.2015.1065414

Table 1

Disentangling Type I Error and Power for the Chi Square and Equivalence based Tests

	Invariance Concluded		**Noninvariance Concluded**	
	TCS	**EQ**	**TCS**	**EQ**
Population Invariance	Correct Decision: Fail to Reject H_0	Power: Correctly Reject H_0	Type I Error: Incorrectly Reject H_0	Type II Error: Fail to reject H_0
Population Noninvariance	Type II Error: Fail to Reject H_0	Type I Error: Incorrectly Reject H_0	Power: Correctly Reject H_0	Correct Decision: Fail to Reject H_0

Note: TCS = traditional chi square (H_0: there is no difference), EQ = equivalence test (H_0: there is a difference as defined by a particular equivalence interval). The grey shading indicates the probability rates of interest in the simulation study. The shaded top row indicates rates of correctly concluding invariance and the shaded bottom row indicates rates of falsely concluding invariance.

Table 2

Applied Example Measurement Invariance Results by Method

MI Stage	**χ^2 statistic**	**EQ**	**EQ-A**	**TCS**	**CFI**	**RMSEA**	**MNCI**
Men	408.79 *df*=80	δ_0 = 625.15, $p < .001$	δ_0 = 725.47, $p < .001$	$p < .001$	.973	.056	.874
Women	364.92 *df*=80	δ_0 = 581.63, $p < .001$	δ_0 = 676.49, $p < .001$	$p < .001$	.971	.058	.882
Configural	773.71 *df*=160	Met (see above)	Met (see above)	$p < .001$	.972	.057	.878
Metric	Δ17.44 *df*=10	δ_0 = 75.42, $p < .001$	δ_0 =105.35, $p < .001$	$p = .065$*	Δ .000	Δ.001	Δ -.006
Scalar	Δ43.60 *df*=10	δ_0 = 75.42, $p = .004$	δ_0 =105.35, $p < .001$	$p < .001$	Δ -.002	Δ .000	Δ -.002
Strict	Δ153.84 *df*=15	δ_0 = 113.14, $p = .88$	δ_0 = 147.79, $p = .37$	NA	Δ -.006	Δ .003	Δ .000*

Note: The *p* values for the equivalence tests are derived from a noncentral chi-square distribution based on δ_0 (see Equation 4). Shaded grey areas show the level of invariance concluded using each of the different statistical approaches. For the equivalence testing methods, Yuan and Chan (2016) recommend assessing model fit in each group separately so configural invariance is satisfied with statistically significant results in both groups.

* denotes that configural invariance would not be satisfied (assuming the original five factor model) using the TCS or MNCI alone.

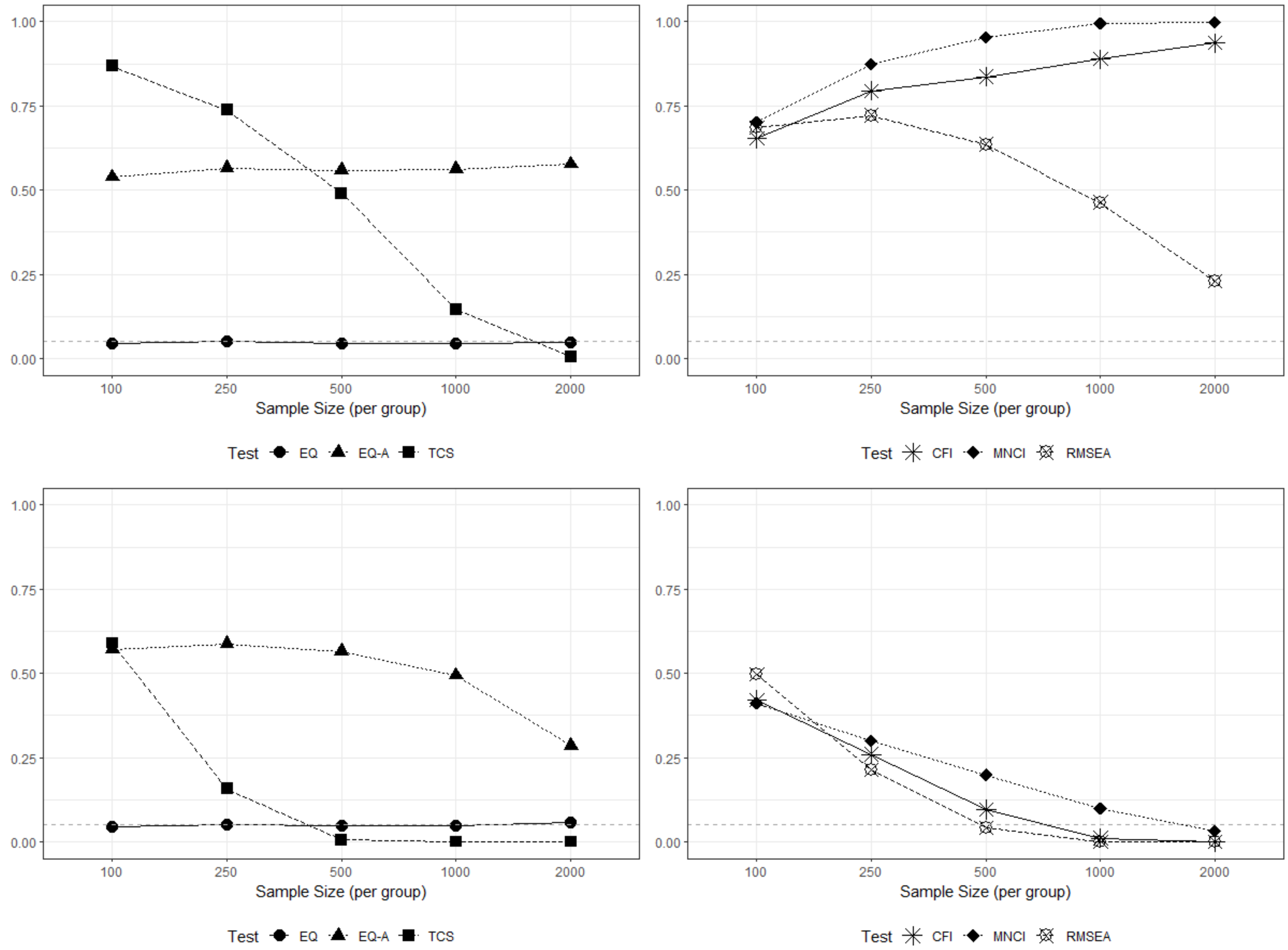

Figure 1. Rates of falsely concluding metric invariance in the 4-indicator model with factor loadings of .7. The top row includes population model misspecification based on RMSEA_0 = .05, whereas the bottom row is based on population model misspecification based on RMSEA_0 = .10.

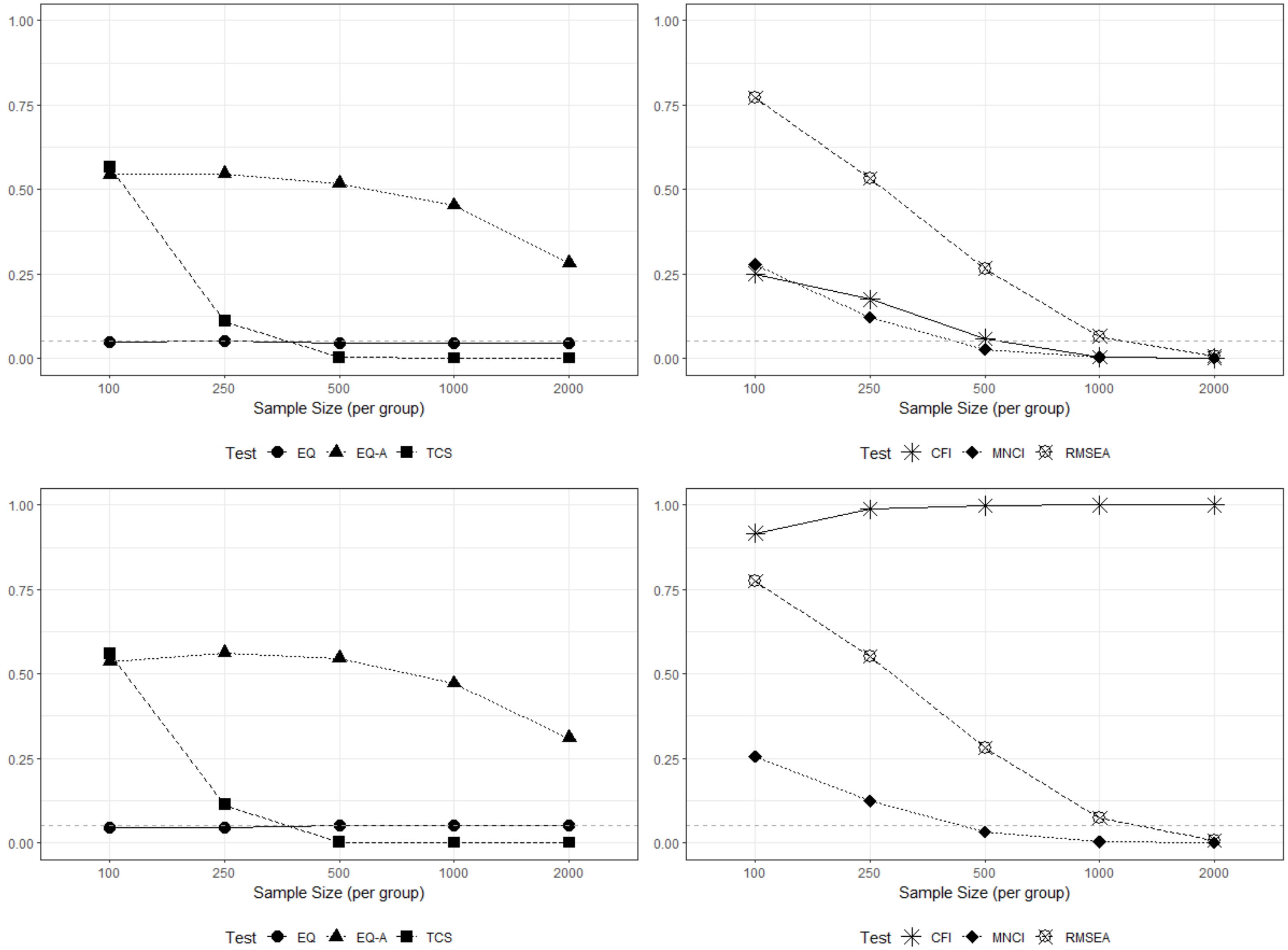

Figure 2. Rates of falsely concluding metric invariance in the eight-indicator model with a population model misspecification of RMSEA$_0$ = .08. The top row includes population factor loadings of .5, whereas the bottom row includes factor loadings at .9. The dotted grey line is at .05, the nominal acceptable rate of falsely concluding invariance.

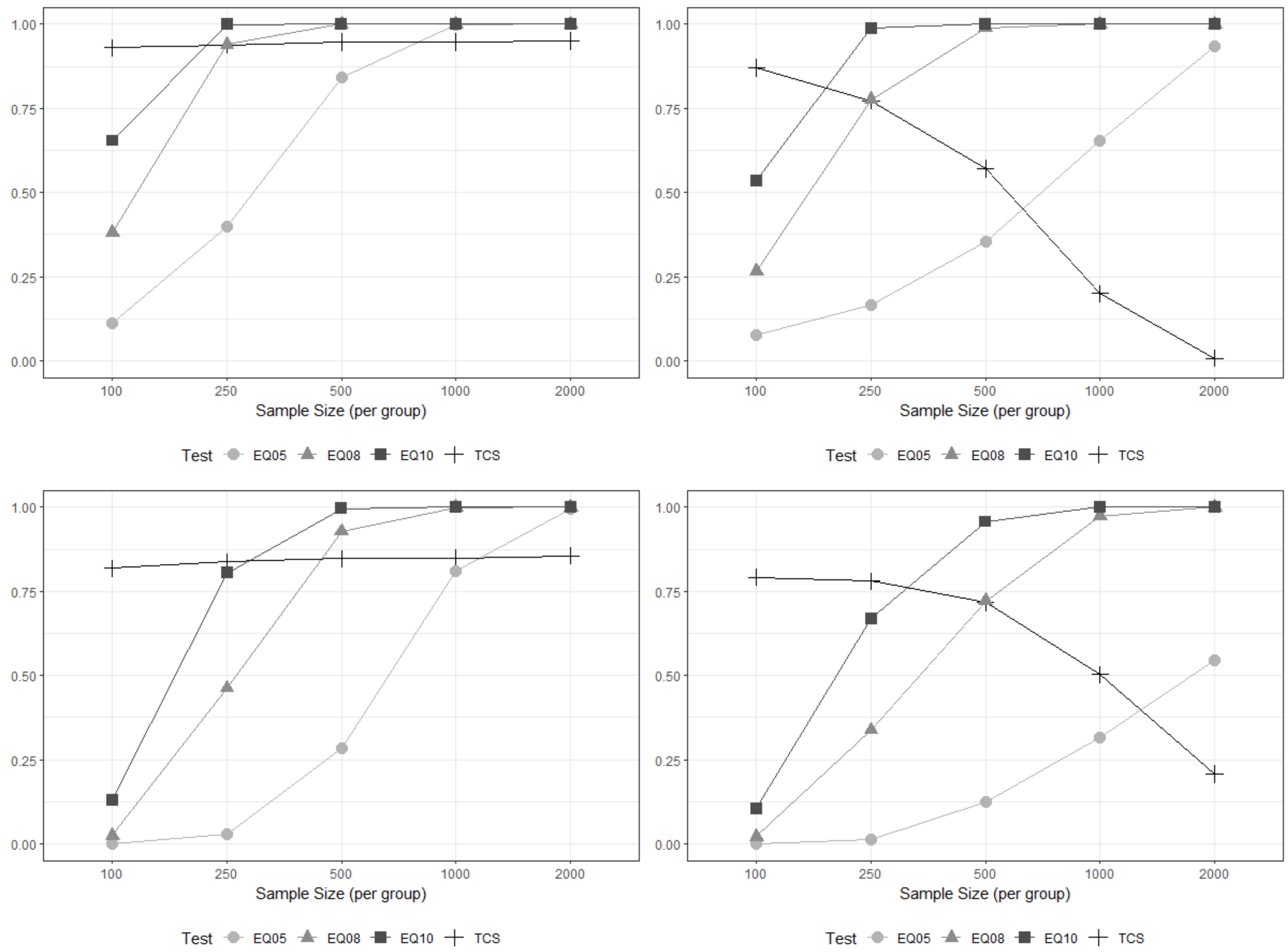

Figure 3. Rates of correctly concluding invariance (power rates for the EQ). The 05, 08, and 10 that follow "EQ" represent different values of RMSEA_0 used for calculating the equivalence interval. The left column represents equal population models and the right column includes results with slightly different population models (with differences still within the equivalence bounds). The top row includes rates of concluding invariance independent of MI sequence to demonstrate the power of the EQ test in and of itself, whereas the bottom row shows metric invariance rates. All results are taken from the four-indicator model with factor loadings of .7.